В. И. Бахтин

# ВВЕДЕНИЕ В ПРИКЛАДНУЮ СТАТИСТИКУ

## Курс лекций

## Часть I

## МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УДК 519.22(075.8)
ББК 22.172я73

Р е ц е н з е н т ы:

доктор физико-математических наук,
профессор *А. Д. Егоров*;
доктор физико-математических наук,
профессор *А. В. Лебедев*

**Бахтин, В. И.**

Б30 Введение в прикладную статистику : курс лекций. В 2 ч. Ч. 1: Математическая статистика / В. И. Бахтин. – Минск: БГУ, 2011. – 91 с.

ISBN 978-985-518-490-5.

В первой части курса лекций изложены основные понятия и теоремы классической математической статистики. Большинство теорем снабжены доказательствами, различные статистические методы иллюстрируются примерами с решениями.

Предназначено для студентов математических специальностей, освоивших курсы теории меры, теории вероятностей и интеграла Лебега.

**УДК 519.22(075.8)**
**ББК 22.172я73**

**ISBN 978-985-518-490-5 (ч. 1)**
**ISBN 978-985-518-491-2**

# ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Мы все часто слышим в экономических блоках новостей слова «статистика», «статистические данные» или даже «национальный статистический комитет». Каждое макроэкономическое исследование или прогноз так или иначе опирается на статистику. Любой опрос общественного мнения — это статистическая выборка. Некоторые статистические данные настолько важны, так сильно могут повлиять на деятельность отдельных предприятий или даже на благосостояние целого государства, что их засекречивают, умышленно искажают, продают за деньги. Поэтому доказывать важность изучения науки, называемой статистикой, нет никакой необходимости.

Однако предмет *математической* статистики несколько отличается от того, что понимает под ней рядовой обыватель. В обыденной жизни слово «статистика» ассоциируется прежде всего со сбором однотипных сведений либо о большом числе объектов, либо на протяжении длительного периода времени, а также с очень поверхностным анализом этих сведений (результаты которого могут выражаться фразами вроде «производство такого-то товара за год выросло на столько-то процентов», «индекс потребительского доверия падает», «такой-то кандидат победит уже в первом туре президентских выборов»). В отличие от этого математическая статистика не занимается сбором сведений. В ней изначально предполагается, что некоторый набор данных уже имеется, а цель — их всесторонний анализ, на основе которого определяют тип и параметры распределения вероятностей, которому подчинены данные, и проверяют различные гипотезы о параметрах этого распределения. В прикладной статистике идут еще дальше: исследуют зависимости между различными переменными, представляют одни из них как функции от других, и на этой основе делают прогнозы (например, как будут варьироваться одни параметры при определенном изменении других параметров, или о развитии событий в будущем).

В общих чертах предмет математической статистики можно описать следующим образом.

Пусть на пространстве $\mathbb{R}^n$ задано семейство распределений вероятностей $P_\theta$, где параметр $\theta$ изменяется в каком-то множестве $Q$.

При некотором значении $\theta \in Q$ рассматривается случайная величина $x \in \mathbb{R}^n$ с распределением $P_\theta$. Для нее проводится $N$ независимых

испытаний, в которых наблюдаются значения $x_1, \ldots, x_N \in \mathbb{R}^n$ (другими словами, все случайные величины $x_1, \ldots, x_N$ независимы и имеют распределение $P_\theta$). Набор $X = (x_1, \ldots, x_N)$ называется *выборкой* объема $N$ из распределения $P_\theta$.

Целью математической статистики является получение информации о параметре $\theta$ по наблюдаемой выборке $X$. При этом считается, что семейство вероятностных распределений $\{ P_\theta \mid \theta \in Q \}$ задано, но истинное значение $\theta$ неизвестно. Эта задача решается с помощью специальным образом подобранных функций от выборки, называемых *статистиками*. Таким образом, слово «статистика» употребляется в двух смыслах: как название научной дисциплины и как некоторая функция от выборки.

В математической статистике выделяются два больших раздела: оценивание параметров и проверка гипотез. В первом из них находят такие статистики $T(X)$, значения которых в некотором вероятностном смысле близки к истинному $\theta$. Эти статистики называются *оценками* $\theta$. Во втором разделе пытаются определить, принадлежит ли $\theta$ некоторым заранее заданным подмножествам в области изменения параметра $Q$. Утверждение о принадлежности $\theta$ какому-либо подмножеству $Q' \subset Q$ называется *гипотезой*.

В данном пособии излагается материал, приблизительно соответствующий семестровому курсу статистики для студентов математических специальностей. Этот материал также служит базой для различных методов прикладной статистики, которые содержатся во второй части курса и предназначены в первую очередь для студентов-математиков, специализирующихся в экономике.

При изучении математической статистики предполагается знакомство с основными понятиями и фактами теории вероятностей и теории интеграла Лебега. Для удобства читателя эти сведения приведены в приложении.

Материал повышенной сложности набран в тексте мелким шрифтом. Названия тех параграфов, материал которых никак не используется в дальнейшем, помечены символом «†».

В заключение этого короткого вступления выражаю глубокую благодарность А. В. Лебедеву, который сделал большое число замечаний, послуживших значительному улучшению рукописи.

# Глава 1. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ

## § 1. Основные понятия статистического оценивания

Начнем с описания базовых объектов, участвующих в процедуре статистического оценивания.

Под *распределением вероятностей* на пространстве $\mathbb{R}^n$ мы будем понимать любую вероятностную борелевскую меру $P$ на $\mathbb{R}^n$ (которая определена на σ-алгебре борелевских множеств, см. Приложение, п. I). Говорят, что случайная величина $x \in \mathbb{R}^n$ имеет распределение $P$, если вероятность ее попадания в любое борелевское множество $A \subset \mathbb{R}^n$ равна $P(A)$. *Выборкой* объема $N$ из распределения $P$ называют последовательность из $N$ независимых случайных величин $X = (x_1, \ldots, x_N)$, каждая из которых имеет распределение $P$. Очевидно, $X \in \mathbb{R}^{nN}$. Если был проведен эксперимент, в результате которого случайные величины $x_1, \ldots, x_N$ приняли какие-то конкретные значения (выборочные значения), то набор этих значений тоже называют выборкой. Поскольку случайные величины $x_1, \ldots, x_N$ независимы, распределение выборки $X$ удовлетворяет тождеству

$$P\{x_1 \in A_1, \ldots, x_N \in A_N\} = P\{x_1 \in A_1\} \times \ldots \times P\{x_N \in A_N\}.$$

В том случае, когда распределение $P$ имеет плотность $p(x)$, распределение выборки тоже имеет плотность $p(X)$, которая представляется как произведение

$$p(X) = p(x_1) \times \ldots \times p(x_N).$$

**Определение.** *Статистика* — это любая (измеримая по Борелю) функция $T(X)$ от выборки $X$. Ее значения могут лежать в любом метрическом пространстве.

Сразу следует отметить, что определяя статистику как функцию от выборки, мы допускаем некоторую вольность речи. Дело в том, что выборка $X = (x_1, \ldots, x_N)$ может иметь какой угодно объем $N$. Так что на самом деле всякая статистика — это целая последовательность

функций от выборок разного объема. Кроме того, все эти функции, как и сами выборки, являются *случайными* величинами.

Пусть на пространстве $\mathbb{R}^n$ задано семейство вероятностных распределений $\{\, P_\theta \mid \theta \in Q \,\}$, где $Q \subset \mathbb{R}^m$, и при некотором (неизвестном для нас) значении параметра $\theta$ наблюдается выборка $X = (x_1, \ldots, x_N)$ из распределения $P_\theta$. Основная задача статистического оценивания заключается в том, чтобы найти статистику $\hat\theta = T(X)$, значение которой было бы в некотором вероятностном смысле близко к истинному значению $\theta$. При этом величина $\hat\theta = T(X)$ называется *статистической оценкой* параметра $\theta$.

**Пример.** Пусть $\mathcal{N}(a, d)$ — нормальное распределение с математическим ожиданием $a$ и дисперсией $d$. Его функция распределения имеет вид

$$F_{ad}(x) = \frac{1}{\sqrt{2\pi d}} \int_{-\infty}^{x} e^{-(u-a)^2/2d}\, du.$$

Предположим, что мы наблюдаем выборку $X = (x_1, \ldots, x_N)$ из этого распределения, а параметры $a$ и $d$ нам неизвестны. Требуется найти для них оценки. В этом примере выборочные значения $x_i$ одномерные, а параметр $\theta = (a, d)$ двумерный. В качестве области $Q$ можно взять полуплоскость $Q = \{\, (a, d) \in \mathbb{R}^2 \mid d > 0 \,\}$.

Оценками для $a$ и $d$ в этом примере могут служить выборочное среднее $\bar x$ и выборочная дисперсия $\hat\sigma^2$, определяемые формулами:

$$\bar x = \frac{1}{N} \sum_{t=1}^{N} x_t, \qquad \hat\sigma^2 = \frac{1}{N} \sum_{t=1}^{N} (x_t - \bar x)^2. \tag{1.1}$$

Таким образом, мы полагаем $\hat\theta = (\bar x, \hat\sigma^2)$. Рисунок 1 иллюстрирует описанную процедуру оценивания.

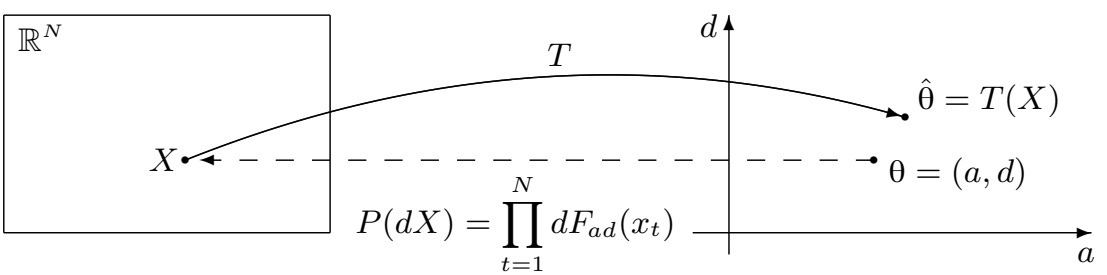


*Рис. 1.* Оценивание параметров нормального распределения

Довольно часто используют более общий подход к понятию параметра распределения вероятностей. А именно, под параметром понимают любой функционал, определенный на множестве *всех* (или значительной части) распределений. Типичными примерами таких функционалов являются математическое ожидание и дисперсия.

Если рассматривается семейство распределений $P_\theta$, зависящее от конечномерного параметра $\theta$, то статистическое оценивание принято называть *параметрическим*. Во втором случае, когда оценивается значение функционала, определенного на множестве всех вероятностных распределений, говорят о *непараметрическом* статистическом оценивании (потому что нет параметрического семейства $P_\theta$). Примеры такого оценивания будут рассмотрены в § 3.

**Определение.** Статистическая оценка $\hat\theta = T(X)$ *состоятельна,* если при любом значении параметра $\theta$ она сходится к $\theta$ по вероятности при $N \to \infty$. В явном виде это означает, что для всякого $\varepsilon > 0$

$$P_\theta\big\{|\hat\theta - \theta| \geqslant \varepsilon\big\} \to 0 \quad \text{при } N \to \infty.$$

Статистическая оценка $\hat\theta = T(X)$ *строго* (или *сильно*) состоятельна, если при каждом $\theta$ она сходится к $\theta$ почти наверное (с вероятностью единица по отношению к распределению $P_\theta$). Здесь предполагается, что $\theta$ и $\hat\theta$ принадлежат евклидову пространству $\mathbb{R}^m$.

По сути, состоятельность оценки означает, что определяемая ею *функциональная* последовательность (как мы уже отмечали, всякая оценка — это последовательность функций на вероятностном пространстве, занумерованная объемом выборки $N$) сходится к константе $\theta$. В курсах теории вероятностей и функционального анализа доказывается, что из сходимости почти наверное (или почти всюду в терминологии теории меры) вытекает сходимость по вероятности (по мере). Поэтому из строгой состоятельности оценки вытекает ее состоятельность.

**Определение.** *Смещением* оценки $\hat\theta = T(X)$ называется математическое ожидание разности $\hat\theta - \theta$:

$$b(\theta) = \mathsf{E}_\theta\big\{\hat\theta - \theta\big\} = \mathsf{E}_\theta\hat\theta - \theta.$$

Если $b(\theta) \equiv 0$, то оценка называется *несмещенной*. Если же $b(\theta) \to 0$ при $N \to \infty$, то оценка называется *асимптотически несмещенной*.

Разумеется, смещение зависит не только от $\theta$, но и от мощности выборки $N$.

Несмещенные оценки существуют не всегда. В качестве примера рассмотрим последовательность испытаний Бернулли $X = (x_1, \ldots, x_N)$, в которой каждое выборочное значение $x_i$ может быть равно единице с вероятностью $\theta^2$ или нулю с вероятностью $1 - \theta^2$, где $\theta \in (0, 1)$. Тогда вероятность всей выборки $X$ вычисляется по формуле

$$P_\theta(X) = \big(\theta^2\big)^{x_1 + \ldots + x_N} \big(1 - \theta^2\big)^{N - x_1 - \ldots - x_N}.$$

Из нее следует, что $P_\theta(X)$ является многочленом от $\theta^2$. Математическое ожидание любой статистики $T(X)$ имеет вид

$$\mathsf{E}_\theta\{T(X)\} = \sum_{X \in \{0,1\}^N} P_\theta(X) T(X)$$

и тоже является многочленом от $\theta^2$. Значит, равенство $\mathsf{E}_\theta\{T(X)\} = \theta$ невозможно.

Тем не менее для многих стандартных семейств распределений, изучаемых в статистике, несмещенные оценки существуют. Например, выборочное среднее $\bar{x}$ всегда является несмещенной оценкой математического ожидания случайной величины $x$. Выборочная дисперсия $\hat{\sigma}^2$ из (1.1) смещена, однако несмещенная оценка существует и равна

$$s^2 = \frac{1}{N-1} \sum_{t=1}^{N} (x_t - \bar{x})^2. \tag{1.2}$$

**Задача 1.** Докажите несмещенность оценки (1.2).

**Задача 2.** Докажите, что если оценка состоятельна и ограничена, то она асимптотически несмещенная.

## § 2. Вариации и ковариации оценок

Мы изучаем семейство вероятностных распределений $\{\, P_\theta \mid \theta \in Q \,\}$ на $\mathbb{R}^n$, где $Q$ — некоторая область в евклидовом пространстве $\mathbb{R}^m$, и статистическую оценку $\hat{\theta}$ для параметра $\theta$. По определению, эта оценка является функцией $\hat{\theta} = T(X)$ от выборки $X = (x_1, \ldots, x_N) \in \mathbb{R}^{nN}$ из распределения $P_\theta$. Очевидно, $\theta$ и $\hat{\theta}$ являются $m$-векторами:

$$\theta = (\theta_1, \ldots, \theta_m), \qquad \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m) = \big(T_1(X), \ldots, T_m(X)\big).$$

Математическим ожиданием оценки $\hat\theta$ называется вектор

$$\mathsf{E}_\theta\hat\theta = \big(\mathsf{E}_\theta\hat\theta_1,\dots,\mathsf{E}_\theta\hat\theta_m\big), \qquad \mathsf{E}_\theta\hat\theta_i = \int_{\mathbb{R}^{nN}} T_i(X)\,P_\theta(dX).$$

Дисперсия (полная дисперсия) оценки $\hat\theta$ — это

$$\mathsf{D}\hat\theta = \mathsf{E}_\theta\Big\{\big|\hat\theta - \mathsf{E}_\theta\hat\theta\big|^2\Big\} = \mathsf{E}_\theta\left\{\sum_{i=1}^m\big(\hat\theta_i - \mathsf{E}_\theta\hat\theta_i\big)^2\right\} = \sum_{i=1}^m \mathsf{D}\hat\theta_i.$$

Для полной дисперсии также верны равенства

$$\mathsf{D}\hat\theta = \sum_{i=1}^m \mathsf{D}\hat\theta_i = \sum_{i=1}^m \mathsf{E}_\theta\hat\theta_i^{\,2} - \sum_{i=1}^m\big(\mathsf{E}_\theta\hat\theta_i\big)^2 = \mathsf{E}_\theta\big|\hat\theta\big|^2 - \big|\mathsf{E}_\theta\hat\theta\big|^2.$$

*Вариацией* (или полной вариацией) оценки $\hat\theta$ называется величина

$$\operatorname{Var}\hat\theta = \mathsf{E}_\theta|\hat\theta-\theta|^2 = \sum_{i=1}^m \mathsf{E}_\theta(\hat\theta_i-\theta_i)^2.$$

Обычно считается, что полная вариация служит мерой уклонения $\hat\theta$ от истинного значения $\theta$. При этом оценка $\hat\theta$ тем лучше, чем меньше ее вариация. Сразу же отметим, что вариацию статистической оценки невозможно сделать сколь угодно малой, потому что для нее всегда существует оценка снизу (см. неравенство Рао — Крамéра в § 6).

*Матрицей вариаций* для $\hat\theta$ называется матрица $V = (v_{ij})_{i,j=1}^m$ с элементами

$$v_{ij} = \mathsf{E}_\theta\big\{(\hat\theta_i-\theta_i)(\hat\theta_j-\theta_j)\big\}.$$

Матрицу вариаций можно также определять формулой

$$V = \mathsf{E}_\theta\big\{(\hat\theta-\theta)\times(\hat\theta-\theta)^*\big\},$$

в которой $(\hat\theta-\theta)$ — вектор-столбец, $(\hat\theta-\theta)^*$ — вектор-строка, а математическое ожидание матрицы $(\hat\theta-\theta)\times(\hat\theta-\theta)^*$ вычисляется поэлементно.

Аналогично, *матрицей ковариаций* оценки $\hat\theta$ называется матрица $\Sigma = (\sigma_{ij})_{i,j=1}^m$ вида

$$\Sigma = \operatorname{Cov}\big\{\hat\theta,\hat\theta\big\} = \mathsf{E}_\theta\big\{\big(\hat\theta-\mathsf{E}_\theta\hat\theta\big)\times\big(\hat\theta-\mathsf{E}_\theta\hat\theta\big)^*\big\},$$

$$\sigma_{ij} = \mathsf{E}_\theta\big\{\big(\hat\theta_i-\mathsf{E}_\theta\hat\theta_i\big)\big(\hat\theta_j-\mathsf{E}_\theta\hat\theta_j\big)\big\}.$$

Очевидно, дисперсия $\mathsf{D}\hat{\theta}$, полная вариация $\operatorname{Var}\hat{\theta}$, а также матрицы $V$ и $\Sigma$ зависят от параметра $\theta$, мощности выборки $N$ и вида функции $\hat{\theta} = T(X)$. Непосредственно из определений вытекает, что

$$\mathsf{D}\hat{\theta} = \sum_{i=1}^{m} \sigma_{ii} = \operatorname{tr}\Sigma, \qquad \operatorname{Var}\hat{\theta} = \sum_{i=1}^{m} v_{ii} = \operatorname{tr} V, \tag{2.1}$$

где $\operatorname{tr}\Sigma$ обозначает след матрицы $\Sigma$. Кроме того, вариации и ковариации обладают следующими свойствами.

**Свойство 2.1.** *Матрицы $V$ и $\Sigma$ симметричны и неотрицательно определены.*

**Свойство 2.2.** *Справедливы равенства*:

$$V = \Sigma + b(\theta)b(\theta)^*, \tag{2.2}$$

$$\operatorname{Var}\hat{\theta} = \mathsf{D}\hat{\theta} + |b(\theta)|^2, \tag{2.3}$$

*где $b(\theta)$ — смещение $\hat{\theta}$. В частности, если оценка $\hat{\theta}$ несмещенная, то тогда $V = \Sigma$.*

**Свойство 2.3.** *Если $\operatorname{Var}\hat{\theta} \to 0$ при $N \to \infty$, то оценка $\hat{\theta}$ состоятельна.*

Доказательство. Симметричность матриц $V$ и $\Sigma$ очевидна. Неотрицательная определенность матрицы $V$ означает, что для любого вектор-столбца $h = (h_1, \ldots, h_m)$ произведение $h^*Vh$ неотрицательно. Это свойство проверяется простой выкладкой

$$h^*Vh = \sum_{i,j=1}^{m} h_i v_{ij} h_j = \mathsf{E}_\theta\left\{\sum_{i,j=1}^{m} h_i(\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)h_j\right\} =$$

$$= \mathsf{E}_\theta\left\{\left(\sum_{i=1}^{m} h_i(\hat{\theta}_i - \theta_i)\right)^2\right\} \geqslant 0.$$

Неотрицательная определенность матрицы $\Sigma$ доказывается точно так же. Равенство (2.2) вытекает из выкладки

$$v_{ij} = \mathsf{E}_\theta\Big\{\Big((\hat{\theta}_i - \mathsf{E}_\theta\hat{\theta}_i) + (\mathsf{E}_\theta\hat{\theta}_i - \theta_i)\Big)\Big((\hat{\theta}_j - \mathsf{E}_\theta\hat{\theta}_j) + (\mathsf{E}_\theta\hat{\theta}_j - \theta_j)\Big)\Big\} =$$

$$= \sigma_{ij} + b_i(\theta)b_j(\theta).$$

Переходя в (2.2) к следам, получаем (2.3). Наконец, свойство 2.3 доказывается с помощью неравенства Чебышёва (см. Приложение, п. I):

$$P_\theta\{|\hat\theta-\theta|\geqslant\varepsilon\}=P_\theta\left\{\frac{|\hat\theta-\theta|^2}{\varepsilon^2}\geqslant 1\right\}\leqslant \mathsf{E}_\theta\left\{\frac{|\hat\theta-\theta|^2}{\varepsilon^2}\right\}=\varepsilon^{-2}\operatorname{Var}\hat\theta\to 0.$$

**Замечание.** Знакоопределенность матриц $V$ и $\Sigma$ является вариантом известного в линейной алгебре утверждения о том, что матрица, составленная из попарных скалярных произведений некоторого набора векторов (матрица Грама), неотрицательно определена.

# § 3. Выборочные оценки

Рассмотрим конечное вероятностное пространство $\Omega_N=\{1,\ldots,N\}$, в котором вероятности элементарных событий $1,\ldots,N$ одинаковы и равны $1/N$. Каждую выборку $X=(x_1,\ldots,x_N)$ со значениями в $\mathbb{R}^n$ (или любом другом метрическом пространстве) мы будем отождествлять со случайной величиной $X:\Omega_N\to\mathbb{R}^n$, которая сопоставляет каждому элементарному событию $t\in\Omega_N$ значение $X(t)=x_t$.

При таком отождествлении всякая выборка $X=(x_1,\ldots,x_N)$ порождает *эмпирическое (выборочное) распределение вероятностей* $P_X$ на пространстве $\mathbb{R}^n$ по следующему правилу: каждому значению $x_t$ приписывается вероятность $1/N$, а вероятность всего остального множества $\mathbb{R}^n\setminus\{x_1,\ldots,x_N\}$ считается равной нулю. Например, математическое ожидание любой функции $f(x)$ по отношению к эмпирическому распределению $P_X$ выглядит как

$$\mathsf{E}_X\{f(x)\}=\frac{1}{N}\sum_{t=1}^{N}f(x_t).$$

В следующем определении формулируется общий принцип построения выборочных оценок для тех параметров, которые являются функционалами на множестве вероятностных распределений (таких, как математическое ожидание, дисперсия, моменты, ковариации, корреляции и т. п.)

**Определение.** Если $\varphi(P)$ — некоторый функционал на множестве вероятностных распределений, то его *выборочной* (*подстановочной*) оценкой называется $\hat\varphi=\varphi(P_X)$ (то есть значение $\varphi$ на эмпирическом распределении, порожденном выборкой $X$).

Здесь предполагается, что функционал $\varphi(P)$ определен либо на множестве всех распределений, либо на некотором его подмножестве, содержащем все эмпирические распределения.

**Примеры.** Рассмотрим любую случайную величину $x \in \mathbb{R}$ с функцией распределения $F(z) = P\{x \leqslant z\}$. В теории вероятностей для нее определяют следующие параметры:

а) математическое ожидание $\mathsf{E}x$;

б) дисперсию $\mathsf{D}x = \mathsf{E}\big\{(x - \mathsf{E}x)^2\big\}$;

в) моменты $\alpha_k = \mathsf{E}\{x^k\}$;

г) характеристическую функцию $\varphi(\lambda) = \mathsf{E}\{e^{i\lambda x}\}$.

Вычислим соответствующие выборочные оценки.

Выборочное математическое ожидание (выборочное среднее) — это

$$\bar{x} = \mathsf{E}_X\{x\} = \frac{1}{N}\sum_{t=1}^{N} x_t.$$

Выборочная дисперсия определяется формулой

$$\hat{\sigma}^2 = \mathsf{E}_X\big\{(x - \bar{x})^2\big\} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2.$$

Выборочный момент $k$-го порядка — это

$$\hat{\alpha}_k = \mathsf{E}_X\big\{x^k\big\} = \frac{1}{N}\sum_{t=1}^{N} x_t^k.$$

Выборочная характеристическая функция — это

$$\hat{\varphi}(\lambda) = \mathsf{E}_X\big\{e^{i\lambda x}\big\} = \frac{1}{N}\sum_{t=1}^{N} e^{i\lambda x_t}.$$

Чтобы было легче строить различные выборочные оценки, очень полезно запомнить такое правило: любая выборочная «характеристика» определяется как та же самая «характеристика» для дискретной случайной величины, порожденной выборкой. Здесь слово «характеристика» представляет собой лингвистическую переменную, принимающую значения «математическое ожидание», «дисперсия», «ковариация», «момент $k$-го порядка» и т. п. Именно по этому правилу были выписаны четыре последние формулы.

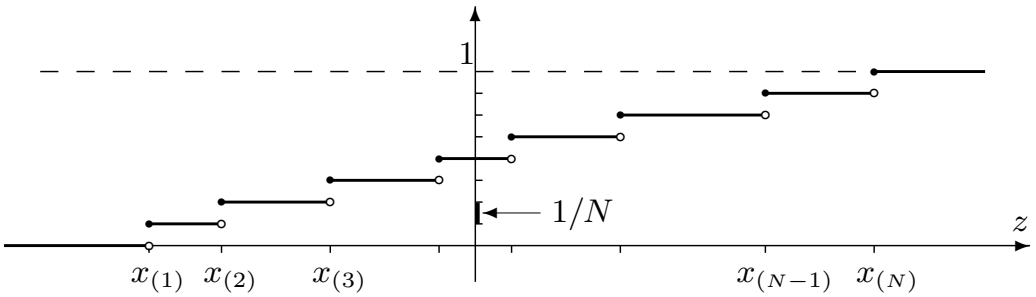


*Рис. 2.* График выборочной функции распределения

В качестве еще одного примера «характеристики», определенной для всех вероятностных распределений на вещественной прямой, рассмотрим функцию распределения $F(z) = P\{x \leqslant z\}$. По нашему правилу ее выборочная оценка $\hat{F}(z)$ определяется как функция распределения, отвечающая эмпирическому распределению вероятностей $P_X$. График выборочной функции распределения $\hat{F}(z)$ изображен на рис. 2. На нем числа $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(N-1)} \leqslant x_{(N)}$ обозначают расположенные в порядке возрастания выборочные значения $x_1, \ldots, x_N$.

Чтобы выписать формулу для $\hat{F}(z)$, введем *функцию Хевисайда*

$$\eta(z) = \begin{cases} 0, & \text{если } \ z < 0, \\ 1, & \text{если } \ z \geqslant 0. \end{cases}$$

Нетрудно видеть, что

$$\hat{F}(z) = \frac{\#\{x_t \mid x_t \leqslant z\}}{N} = \frac{1}{N} \sum_{t=1}^{N} \eta(z - x_t). \tag{3.1}$$

Здесь символ # («диез») обозначает число элементов множества.

В следующей теореме доказывается состоятельность большинства выборочных оценок, перечисленных выше.

**Теорема 3.1.** *Для любых фиксированных* $z, \lambda \in \mathbb{R}$ *статистики* $\hat{F}(z)$, $\hat{\alpha}_k$ *и* $\hat{\varphi}(\lambda)$ *являются сильно состоятельными и несмещенными оценками соответственно для* $F(z)$, $\alpha_k$ *и* $\varphi(\lambda)$. *При этом вариации* $\hat{F}(z)$ *и* $\hat{\alpha}_k$ *имеют вид*

$$\operatorname{Var} \hat{F}(z) = \frac{F(z)(1 - F(z))}{N}, \qquad \operatorname{Var} \hat{\alpha}_k = \frac{\alpha_{2k} - \alpha_k^2}{N}$$

(*последняя — при условии, что существует момент* $\alpha_{2k}$).

Доказательство. Положим $\zeta_t = \eta(z - x_t)$. Тогда

$$P\{\zeta_t = 1\} = P\{x_t \leqslant z\} = F(z),$$

$$P\{\zeta_t = 0\} = P\{x_t > z\} = 1 - F(z),$$

$$\mathsf{E}\zeta_t = 1 \cdot F(z) + 0 \cdot (1 - F(z)) = F(z),$$

$$\mathsf{D}\zeta_t = \mathsf{E}\zeta_t^2 - (\mathsf{E}\zeta_t)^2 = F(z)(1 - F(z)).$$

По определению случайные величины $\zeta_1, \ldots, \zeta_N$ независимы. Значит, они образуют последовательность испытаний Бернулли. В силу (3.1)

$$\hat{F}(z) = \frac{1}{N}\sum_{t=1}^{N} \eta(z - x_t) = \frac{1}{N}\sum_{t=1}^{N} \zeta_t.$$

По усиленному закону больших чисел последняя сумма с вероятностью один сходится к $\mathsf{E}\zeta_t = F(z)$. Значит, оценка $\hat{F}(z)$ для $F(z)$ сильно состоятельна. Она несмещенная, поскольку $\mathsf{E}\hat{F}(z) = \mathsf{E}\zeta_t = F(z)$. Вычислим ее вариацию:

$$\operatorname{Var}\hat{F}(z) = \mathsf{D}\hat{F}(z) = \frac{1}{N^2}\sum_{t=1}^{N} \mathsf{D}\zeta_t = \frac{F(z)(1 - F(z))}{N}.$$

Несмещенность и сильная состоятельность оценок $\hat{\alpha}_k$ и $\hat{\varphi}(\lambda)$ доказывается аналогично. Найдем вариацию $\alpha_k$:

$$\operatorname{Var}\hat{\alpha}_k = \mathsf{D}\hat{\alpha}_k = \mathsf{D}\left\{\frac{1}{N}\sum_{t=1}^{N} x_t^k\right\} = \frac{1}{N}\mathsf{D}\left\{x_t^k\right\} =$$

$$= \frac{1}{N}\left(\mathsf{E}x_t^{2k} - \left(\mathsf{E}x_t^k\right)^2\right) = \frac{\alpha_{2k} - \alpha_k^2}{N}. \quad \square$$

**Вопрос 1.** Почему в этой теореме не вычисляется вариация $\hat{\varphi}(\lambda)$?

**Вопрос 2.** Вытекает ли из этой теоремы сильная состоятельность и несмещенность выборочной дисперсии?

**Задача 1.** При любых фиксированных значениях $z$ и $\lambda$ отображения $X \mapsto \hat{F}(z)$ и $X \mapsto \hat{\varphi}(\lambda)$ являются скалярными статистиками. Мы только что доказали, что они сильно состоятельны. Но можно также

рассматривать и функциональные статистики $X \mapsto \hat{F}(\,\cdot\,)$ и $X \mapsto \hat{\varphi}(\,\cdot\,)$, принимающие значения в множестве функций на вещественной оси. Докажите, что они тоже сильно состоятельны (сходимость в функциональном пространстве понимать в смысле равномерной нормы). Это утверждение называется теоремой Гливенко — Кантелли.

**Задача 2.** Докажите, что если параметр $\theta$ является непрерывным функционалом на множестве всех вероятностных распределений, то его выборочная оценка $\hat{\theta}$ сильно состоятельна. Будут ли математическое ожидание и дисперсия непрерывными функционалами на множестве вероятностных распределений? Как определяется топология на множестве вероятностных распределений? (См. Приложение, п. IV).

**Теорема 3.2.** *Если* $0 < F(z) < 1$, *то для всех* $u \in \mathbb{R}$

$$P\left\{\sqrt{N}\frac{\hat{F}(z) - F(z)}{\sqrt{F(z)(1-F(z))}} \leqslant u\right\} \longrightarrow \Phi(u) \quad \textit{при} \quad N \to \infty,$$

*а если существует момент* $\alpha_{2k}$ *и при этом* $\alpha_k^2 < \alpha_{2k}$, *то*

$$P\left\{\sqrt{N}\frac{\hat{\alpha}_k - \alpha_k}{\sqrt{\alpha_{2k} - \alpha_k^2}} \leqslant u\right\} \longrightarrow \Phi(u) \quad \textit{при} \quad N \to \infty,$$

*где* $\Phi(u)$ — *стандартная нормальная функция распределения*:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-x^2/2}\, dx.$$

Поэтому говорят, что оценки $\hat{F}(z)$ и $\hat{\alpha}_k$ асимптотически нормально распределены.

Доказательство. Это вытекает из предыдущей теоремы и центральной предельной теоремы для последовательности независимых одинаково распределенных случайных величин. $\square$

**Замечание.** Условие $\alpha_k^2 < \alpha_{2k}$ как правило выполняется. Это следует из того, что

$$\alpha_{2k} - \alpha_k^2 = \mathsf{E}x_t^{2k} - \left(\mathsf{E}x_t^k\right)^2 = \mathsf{D}x_t^k \geqslant 0.$$

Выборочная функция распределения разрывна, и у нее нет плотности. Вместо нее для выборки $X = (x_1, \ldots, x_N)$ обычно строят так

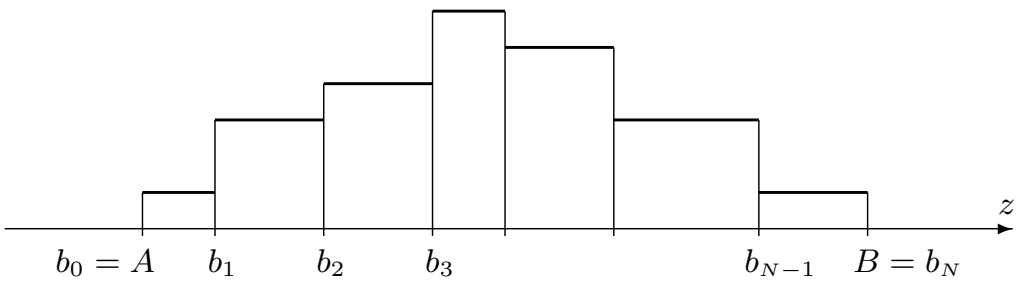


*Рис. 3.* Гистограмма

называемую *гистограмму*. Пусть все выборочные значения $x_t$ лежат на отрезке $[A, B]$. Разобьем его на $n$ частей точками $b_0 < b_1 < \cdots < b_n$. Введем обозначения $\Delta_k = [b_{k-1}, b_k)$ и $\nu_k = \#\{x_t \mid x_t \in \Delta_k\}$. Тогда на каждом интервале $\Delta_k$ полагают $\hat{f}(z) = \nu_k / N|\Delta_k|$. Построенная таким образом гистограмма (см. рис. 3) является плотностью вероятности, потому что интеграл от нее по всей вещественной оси равен единице.

Отметим, что если число интервалов разбиения $n$ фиксировано, то в большинстве случаев гистограмма является смещенной и несостоятельной оценкой для плотности распределения $f(z) = F'(z)$ (уже хотя бы по той простой причине, что функцию $f(z)$ невозможно аппроксимировать ступенчатыми функциями с $n$ значениями).

## § 4. Квантили и $p$-уровни

Пусть задана скалярная случайная величина $x$ с функцией распределения $F(z) = P\{x \leqslant z\}$. *Квантилью* уровня $p \in (0, 1)$ называется число

$$z_p = \sup\{\, z \mid F(z) < p \,\} = \inf\{\, z \mid F(z) \geqslant p \,\}. \tag{4.1}$$

В частности, для $p = 1/2$ квантиль $z_{1/2}$ называется *медианой*. Известно, что функция распределения всегда непрерывна справа. Отсюда вытекает, что $F(z_p) \geqslant p$. Если к тому же функция распределения непрерывна в точке $z_p$, то тогда $F(z_p) = p$. А в случае, когда функция $F(z)$ непрерывна и строго возрастает на всей вещественной оси, имеет место равенство $z_p = F^{-1}(p)$ (см. рис. 4).

Напомним, что по определению случайная величина $x$ является функцией на каком-то вероятностном пространстве $x\colon \Omega \to \mathbb{R}$. Определим новую случайную величину $F(x)$, являющуюся результатом под-

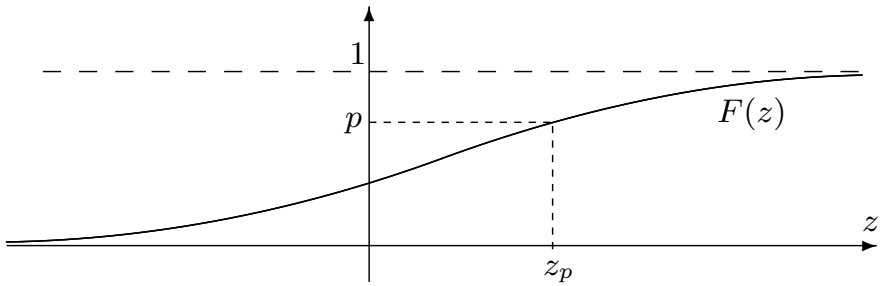


*Рис. 4.* Квантиль уровня $p$

становки случайной величины $x$ в качестве аргумента ее функции распределения $F$ (другими словами, композицию двух функций $x\colon \Omega \to \mathbb{R}$ и $F\colon \mathbb{R} \to [0,1]$). Полученную таким образом случайную величину $F(x)$ называют *p-уровнем* исходной случайной величины $x$.

**Теорема 4.1.** *Если функция распределения случайной величины $x$ непрерывна на вещественной оси, то ее p-уровень $F(x)$ равномерно распределен на отрезке* $[0,1]$.

Доказательство. Действительно, для любого $p \in (0,1)$ из определения квантили и непрерывности распределения следует, что

$$P\{F(x) < p\} = P\{x < z_p\} = P\{x \leqslant z_p\} = F(z_p) = p. \quad \square$$

**Теорема 4.2.** *Для любой случайной величины $x \in \mathbb{R}$ и $p \in [0,1]$*

$$P\big\{F(x-0) < p\big\} \geqslant p, \qquad P\big\{F(x) > p\big\} \geqslant 1-p. \tag{4.2}$$

Доказательство. Пусть $p \in (0,1)$. Если $F(z_p - 0) = p$, то

$$P\big\{F(x-0) < p\big\} = P\{x < z_p\} = F(z_p - 0) = p.$$

А если $F(z_p - 0) < p$, то

$$P\big\{F(x-0) < p\big\} = P\{x \leqslant z_p\} = F(z_p) \geqslant p.$$

Тем самым мы доказали левое из неравенств (4.2) для всех $p \in (0,1)$. В случае $p = 1$ оно получается предельным переходом при $p \to 1-0$, а в случае $p = 0$ оно очевидно.

Пусть $G(z)$ — функция распределения случайной величины $-x$. Для нее выполняется равенство

$$F(z) + G(-z-0) = P\{x \leqslant z\} + P\{-x < -z\} = 1.$$

Из него с помощью уже доказанного левого неравенства (4.2) получаем

$$P\{F(x) > p\} = P\{G(-x-0) < 1-p\} \geqslant 1-p. \quad \square$$

Рассмотрим выборку $X = (x_1, \ldots, x_N)$ из распределения $F(z)$. Расположим все выборочные значения $x_i$ в порядке возрастания:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ \ldots \ \leqslant x_{(N-1)} \leqslant x_{(N)}. \tag{4.3}$$

Случайные величины $x_{(i)}$ называются *порядковыми статистиками*, а вся упорядоченная последовательность (4.3) называется *вариационным рядом*.

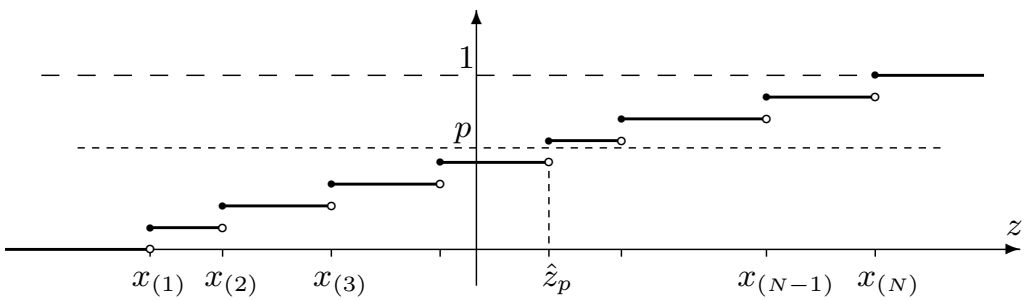


*Рис. 5.* Выборочная квантиль уровня $p$

Всякая выборка $X$ порождает свою выборочную функцию распределения $\hat{F}(z)$. В соответствии с общим правилом, выборочной квантилью уровня $p$ называется $p$-квантиль $\hat{z}_p$ для выборочной функции распределения $\hat{F}(z)$ (см. рис. 5). Она удовлетворяет условиям

$$\hat{z}_p = x_{(i)}, \quad \text{если} \quad \frac{i-1}{N} < p \leqslant \frac{i}{N},$$

и

$$p \leqslant \hat{F}(\hat{z}_p) < p + \frac{1}{N}. \tag{4.4}$$

Традиционно *выборочной медианой* называют не выборочную квантиль $\hat{z}_{1/2}$, как можно было бы ожидать, а центральное значение вариационного ряда, определяемое равенством

$$\hat{\zeta} = \begin{cases} x_{(m)} & \text{при} \quad N = 2m-1, \\ \dfrac{x_{(m)} + x_{(m+1)}}{2} & \text{при} \quad N = 2m. \end{cases}$$

В следующей теореме утверждается, что выборочная квантиль, как и большинство других выборочных оценок, состоятельна.

**Теорема 4.3.** *Если $F(z) > p$ при всех $z > z_p$, то оценка $\hat{z}_p$ для $z_p$ состоятельна.*

Доказательство. Фиксируем произвольное число $\varepsilon > 0$. Тогда по условию теоремы $F(z_p + \varepsilon) > p$. С другой стороны, по определению квантили $F(z_p - \varepsilon) < p$. Выберем такое малое число $\delta > 0$, чтобы одновременно было $F(z_p + \varepsilon) > p + \delta$ и $F(z_p - \varepsilon) < p - \delta$. Из сильной состоятельности выборочной функции распределения вытекает, что

$$\hat{F}(z_p + \varepsilon) \xrightarrow{\text{п. н.}} F(z_p + \varepsilon), \qquad \hat{F}(z_p - \varepsilon) \xrightarrow{\text{п. н.}} F(z_p - \varepsilon)$$

при $N \to \infty$. Следовательно, с вероятностью 1 при всех достаточно больших $N$ выполняются неравенства

$$\hat{F}(z_p + \varepsilon) > p + \delta, \qquad \hat{F}(z_p - \varepsilon) < p - \delta. \tag{4.5}$$

Из (4.4) и (4.5) получаем, что с вероятностью 1

$$z_p - \varepsilon < \hat{z}_p < z_p + \varepsilon \quad \text{при} \quad N \to \infty.$$

В силу произвольности $\varepsilon$ отсюда вытекает, что $\hat{z}_p$ с вероятностью 1 сходится к $z_p$. $\square$

Без доказательства приведем еще одну теорему.

**Теорема 4.4** [3]. *Если функция распределения $F(z)$ имеет плотность $f(z)$, и в некоторой окрестности квантили $z_p$ эта плотность непрерывно дифференцируема и положительна, то выборочная квантиль $\hat{z}_p$ асимптотически нормально распределена: при $N \to \infty$*

$$P\left\{\sqrt{N}\,\frac{\hat{z}_p - z_p}{\sqrt{p(1-p)}}\,f(z_p) \leqslant u\right\} \longrightarrow \Phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-x^2/2}\,dx.$$

# § 5. Метод моментов

Метод моментов для построения статистических оценок был предложен в 1900 г. английским статистиком К. Пирсоном.

Пусть на вещественной оси задано семейство распределений вероятностей $P_\theta$, зависящее от параметра $\theta \in \mathbb{R}^m$, и существуют моменты

$$\alpha_k(\theta) = \mathsf{E}_\theta\{x^k\} = \int_{-\infty}^{\infty} x^k\,P_\theta(dx), \qquad k = 1, \ldots, m.$$

По выборке $X = (x_1, \ldots, x_N)$ из распределения $P_\theta$ найдем выборочные моменты

$$\hat{\alpha}_k = \frac{1}{N}\sum_{t=1}^{N} x_t^k, \qquad k = 1, \ldots, m.$$

**Определение.** Оценкой по методу моментов $\hat{\theta}$ называют решение системы уравнений

$$\alpha_k\big(\hat{\theta}\big) = \hat{\alpha}_k, \qquad k = 1, \ldots, m. \tag{5.1}$$

Вообще говоря, решение у этой системы не всегда существует и не всегда единственно. Для его существования и единственности необходима обратимость отображения $\alpha(\theta) = (\alpha_1(\theta), \ldots, \alpha_m(\theta))$ на области значений отображения $\hat{\alpha}(X) = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)$.

**Пример.** Рассмотрим семейство $\mathcal{N}(a, \sigma^2)$ нормальных распределений с неизвестными математическим ожиданием $a$ и дисперсией $\sigma^2$. Для него первые два момента имеют вид

$$\alpha_1 = \mathsf{E}x = a, \qquad \alpha_2 = \mathsf{E}\big\{x^2\big\} = \mathsf{D}x + a^2 = \sigma^2 + a^2.$$

Приравняем их к выборочным моментам:

$$a = \frac{1}{N}\sum_{t=1}^{N} x_t, \qquad \sigma^2 + a^2 = \frac{1}{N}\sum_{t=1}^{N} x_t^2.$$

Решая эту систему относительно $a$ и $\sigma^2$, получаем оценки $\hat{a} = \bar{x}$ и $\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N} x_t^2 - \bar{x}^2$.

**Теорема 5.1.** *Если для семейства распределений $P_\theta$, где $\theta$ изменяется в открытой области $Q \subset \mathbb{R}^m$, существуют моменты $\alpha_k(\theta)$, $k = 1, \ldots, m$, и отображение $\alpha(\theta) = (\alpha_1(\theta), \ldots, \alpha_m(\theta))$ задает гомеоморфизм области $Q$ на $\alpha(Q)$, то оценка по методу моментов сильно состоятельна.*

Доказательство. По теореме 3.1 оценки $\hat{\alpha}_k$ сходятся к $\alpha_k(\theta)$ с вероятностью 1. Оценка по методу моментов $\hat{\theta}$ — это решение уравнения $\alpha(\hat{\theta}) = \hat{\alpha}(X)$. Оно имеет вид $\hat{\theta} = \alpha^{-1}\big(\hat{\alpha}(X)\big)$ и почти наверное сходится к $\alpha^{-1}(\alpha(\theta)) = \theta$. $\square$

Можно доказать, что если отображение $\alpha(\theta)$ является диффеоморфизмом (иначе говоря, если оно само и обратное к нему отображение непрерывно дифференцируемы), то оценка по методу моментов асимп-

тотически нормально распределена.

Иногда применяют обобщенный метод моментов. Он состоит в том, что для некоторых борелевских функций $g_1(x), \dots, g_m(x)$ определяют *обобщенные моменты* $\alpha_k(\theta)$ и $\hat{\alpha}_k$ по формулам:

$$\alpha_k(\theta) = \int_{-\infty}^{\infty} g_k(x)\, P_\theta(dx), \qquad \hat{\alpha}_k = \frac{1}{N}\sum_{t=1}^{N} g_k(x_t),$$

и находят оценку $\hat{\theta}$ как решение системы (5.1). Для нее теорема 5.1 остается в силе. Оценки, полученные методом обобщенных моментов, называют *подстановочными*.

Одним из примеров применения метода обобщенных моментов служит *метод наименьших квадратов,* который будет изложен во второй части курса.

**Задача.** Докажите теорему 5.1 для обобщенных моментов.

## § 6. Неравенство Рао — Крамéра

Неравенство Рао — Крамéра, называемое также *неравенством информации,* устанавливает нижнюю границу для вариации статистической оценки $\operatorname{Var}\hat{\theta} = \mathsf{E}_\theta\{(\hat{\theta}-\theta)^2\}$. Оказывается, эта граница имеет порядок $1/N$ (где $N$ — объем выборки), и никак не может быть уменьшена. Таким образом, всякая статистическая оценка имеет вполне определенный неулучшаемый предел точности.

В статистике обычно рассматривают два типа вероятностных распределений на пространстве $\mathbb{R}^n$. Распределения первого типа имеют борелевскую плотность по отношению к мере Лебега и называются *непрерывными*. Распределения второго типа сосредоточены на конечном или счетном множестве точек и называются *дискретными*. Чтобы рассматривать семейства непрерывных и дискретных распределений с единой точки зрения, введем понятие *считающей меры*. По определению, считающая мера сосредоточена на конечном или счетном подмножестве $\Omega = \{a_1, a_2, \dots\} \subset \mathbb{R}^n$; при этом $\mu(a_k) = 1$ для любого $a_k \in \Omega$ и $\mu(\mathbb{R}^n \setminus \Omega) = 0$.

Ниже буква $\mu$ у нас будет обозначать либо меру Лебега, либо какую-нибудь считающую меру, либо произвольную $\sigma$-конечную борелевскую меру на $\mathbb{R}^n$.

Пусть на пространстве $\mathbb{R}^n$ задано семейство борелевских плотностей вероятностей $p_\theta(x)$ по отношению к мере $\mu$ (которая в такой ситуации называется *доминирующей*). Иначе говоря, для любого борелевского множества $A \subset \mathbb{R}^n$

$$P_\theta(A) = \int_A p_\theta(x)\,\mu(dx).$$

Тогда математическое ожидание любой случайной величины $f = f(x)$, являющейся борелевской функцией на $\mathbb{R}^n$, вычисляется по формуле

$$\mathsf{E}_\theta f = \int_{\mathbb{R}^n} f(x)p_\theta(x)\,\mu(dx).$$

В случае считающей меры $\mu$, отвечающей конечному или счетному подмножеству $\Omega \subset \mathbb{R}^n$, последние две формулы принимают вид

$$P_\theta(A) = \sum_{a_k \in A\cap\Omega} p_\theta(a_k), \qquad \mathsf{E}_\theta f = \sum_{a_k\in\Omega} f(a_k)p_\theta(a_k).$$

Первая из этих формул означает, что $p_\theta(a_k)$ совпадает с вероятностью события $a_k$.

Рассмотрим выборку $X = (x_1, \ldots, x_N) \in \mathbb{R}^{nN}$ из распределения с плотностью $p_\theta$. Она имеет плотность $p_\theta(X) = p_\theta(x_1)\ldots p_\theta(x_N)$ по отношению к мере $\mu^N$, где $\mu^N(dX) = \mu(dx_1)\ldots\mu(dx_N)$. Следовательно, для любой статистики $T(X)$

$$\mathsf{E}_\theta\{T(X)\} = \int_{\mathbb{R}^{nN}} T(X)p_\theta(X)\,\mu^N(dX).$$

Предположим, что $\theta \in \mathbb{R}$, и что все плотности $p_\theta(x)$ строго положительны и дифференцируемы по $\theta$. Определим функции:

$$\rho_\theta(x) = \frac{p'_\theta(x)}{p_\theta(x)} = \frac{d\ln p_\theta(x)}{d\theta}, \qquad \rho_\theta(X) = \frac{p'_\theta(X)}{p_\theta(X)} = \frac{d\ln p_\theta(X)}{d\theta}; \tag{6.1}$$

$$\mathcal{I}(\theta) = \mathsf{E}_\theta\big\{\rho^2_\theta(x)\big\}, \qquad \mathcal{I}_N(\theta) = \mathsf{E}_\theta\big\{\rho^2_\theta(X)\big\}. \tag{6.2}$$

Функции $\mathcal{I}(\theta)$ и $\mathcal{I}_N(\theta)$ принято называть *количеством информации* (по Фишеру).

Стоит отметить, что информация, в том числе и только что введенная информация по Фишеру, является таким же фундаментальным физическим понятием, как и масса. Несмотря на это, физическая сущность информации пока еще недостаточно осознана и не до конца изучена.

**Лемма 6.1.** *Если выражение* $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx)$ *можно дифференцировать по* $\theta$ *под знаком интеграла, то* $\mathsf{E}_\theta\{\rho_\theta(X)\} = 0$. *Если к тому же функция* $\mathcal{I}(\theta)$ *конечна, то* $\mathcal{I}_N(\theta) = N\mathcal{I}(\theta)$.

Доказательство. Поскольку $p_\theta(x)$ является плотностью вероятности, для нее выполняется равенство $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx) = 1$. Продифференцируем его по $\theta$:

$$\frac{d}{d\theta}\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx) = \int_{\mathbb{R}^n} \frac{dp_\theta(x)}{d\theta}\,\mu(dx) = \int_{\mathbb{R}^n} \frac{p'_\theta(x)}{p_\theta(x)} p_\theta(x)\,\mu(dx) =$$

$$= \mathsf{E}_\theta\{\rho_\theta(x)\} = 0.$$

Из (6.1) и равенства $p_\theta(X) = p_\theta(x_1)\dots p_\theta(x_N)$ вытекает, что

$$\rho_\theta(X) = \rho_\theta(x_1) + \ \dots\ + \rho_\theta(x_N).$$

Следовательно, $\mathsf{E}_\theta\{\rho_\theta(X)\} = 0$.

Далее, учитывая независимость выборочных значений $x_1, \dots, x_N$, получаем

$$\mathcal{I}_N(\theta) = \mathsf{E}_\theta\big\{\rho_\theta^2(X)\big\} =$$

$$= \mathsf{E}_\theta\left\{\left(\sum_{t=1}^N \rho_\theta(x_t)\right)^2\right\} = \sum_{t=1}^N \mathsf{E}_\theta\big\{\rho_\theta^2(x_t)\big\} + \sum_{t\neq s} \mathsf{E}_\theta\big\{\rho_\theta(x_t)\rho_\theta(x_s)\big\} =$$

$$= N\mathsf{E}_\theta\big\{\rho_\theta^2(x)\big\} + \sum_{t\neq s} \mathsf{E}_\theta\{\rho_\theta(x_t)\}\mathsf{E}_\theta\{\rho_\theta(x_s)\} = N\mathcal{I}(\theta). \quad \square$$

**Теорема 6.2 (неравенство Рао — Крамéра).** *Пусть на* $\mathbb{R}^n$ *задана* $\sigma$*-конечная мера* $\mu$, *имеется семейство всюду положительных плотностей вероятности* $\{p_\theta(x) \mid \theta \in (a,b)\}$ *по отношению к* $\mu$, *и пусть для параметра* $\theta$ *есть статистическая оценка* $\hat\theta = T(X)$. *Если выражения* $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx)$ *и* $\int_{\mathbb{R}^{nN}} T(X)p_\theta(X)\,\mu^N(dX)$ *можно дифференцировать по* $\theta$ *под знаком интеграла, то*

$$\operatorname{Var}\hat\theta = \mathsf{E}_\theta\Big\{\big(\hat\theta - \theta\big)^2\Big\} \geqslant \frac{(1+b'(\theta))^2}{N\mathcal{I}(\theta)} + b^2(\theta), \tag{6.3}$$

*где* $b(\theta) = \mathsf{E}_\theta\big\{\hat\theta - \theta\big\}$ — *смещение* $\hat\theta$. *Если оценка* $\hat\theta$ *несмещенная, то*

$$\operatorname{Var}\hat\theta = \mathsf{E}_\theta\Big\{\big(\hat\theta - \theta\big)^2\Big\} \geqslant \frac{1}{N\mathcal{I}(\theta)}. \tag{6.4}$$

Доказательство. По определению смещения

$$\mathsf{E}_\theta\hat\theta = \int_{\mathbb{R}^{nN}} T(X)p_\theta(X)\,\mu^N(dX) = \theta + b(\theta).$$

Продифференцируем последнее равенство по $\theta$:

$$1 + b'(\theta) = \int_{\mathbb{R}^{nN}} T(X)\frac{p'_\theta(X)}{p_\theta(X)}p_\theta(X)\,\mu^N(dX) = \mathsf{E}_\theta\big\{\hat\theta\rho_\theta(X)\big\}. \tag{6.5}$$

Используя (6.5) и доказанное в лемме 6.1 тождество $\mathsf{E}_\theta\{\rho_\theta(X)\} = 0$, получаем

$$\mathsf{E}_\theta\Big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta - c\rho_\theta(X)\big)^2\Big\} =$$

$$= \mathsf{E}_\theta\Big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta\big)^2\Big\} - 2c\mathsf{E}_\theta\big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta\big)\rho_\theta(X)\big\} + c^2\mathsf{E}_\theta\big\{\rho_\theta^2(X)\big\} =$$

$$= \mathsf{E}_\theta\Big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta\big)^2\Big\} - 2c(1 + b'(\theta)) + c^2\mathcal{I}_N(\theta) \geqslant 0, \tag{6.6}$$

где $c$ — произвольная константа. Последнее выражение принимает свое минимальное значение при $c = (1 + b'(\theta))/\mathcal{I}_N(\theta)$. В этом случае неравенство (6.6) превращается в

$$\mathsf{E}_\theta\Big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta\big)^2\Big\} \geqslant \frac{(1 + b'(\theta))^2}{\mathcal{I}_N(\theta)}. \tag{6.7}$$

Напомним, что в силу свойств вариаций (см. формулу (2.3))

$$\operatorname{Var}\hat\theta = \mathsf{D}\hat\theta + b^2(\theta) = \mathsf{E}_\theta\Big\{\big(\hat\theta - \mathsf{E}_\theta\hat\theta\big)^2\Big\} + b^2(\theta). \tag{6.8}$$

Объединяя (6.7) и (6.8), получаем (6.3). □

Поскольку $\mathsf{E}_\theta\{\rho_\theta(X)\} = 0$, количество информации $N\mathcal{I}(\theta) = \mathsf{E}_\theta\{\rho_\theta^2(X)\}$ совпадает с дисперсией $\mathsf{D}\{\rho_\theta(X)\}$. Поэтому неравенство информации (6.4) можно интерпретировать как принцип неопределенности

$$\operatorname{Var}\hat\theta \cdot \mathsf{D}\{\rho_\theta(X)\} \geqslant 1.$$

Он тесным образом связан с принципом неопределенности Гейзенберга в квантовой механике.

Обычно в учебниках по статистике на семейство плотностей $p_\theta(x)$ и оценку $\hat\theta = T(X)$ накладываются те или иные условия регулярности, обеспечивающие дифференцируемость по параметру интегралов

из теоремы 6.2. Для простоты изложения мы их не формулируем. На практике условия теоремы 6.2 выполняются не для всех семейств распределений. Например, они нарушаются, если плотности $p_\theta(x)$ при разных $\theta$ имеют разные носители (множества, на которых эти плотности отличны от нуля). В таком случае вариация оценки может оказаться по порядку меньше $1/N$. Этот эффект иллюстрирует следующая задача.

**Задача 1.** Докажите, что для семейства равномерных распределений на отрезке $[0, \theta]$ оценка $T(X) = \max_{1 \leqslant i \leqslant N} x_i$ для параметра $\theta$ имеет вариацию порядка $1/N^2$.

В случае многомерного параметра $\theta = (\theta_1, \ldots, \theta_m)$ определяют *информационную матрицу* $\mathcal{I}(\theta)$ размерности $m \times m$ с элементами

$$\mathcal{I}_{kl}(\theta) = \mathsf{E}_\theta \left\{ \frac{\partial \ln p_\theta(x)}{\partial \theta_k} \, \frac{\partial \ln p_\theta(x)}{\partial \theta_l} \right\}.$$

В матричной форме записи $\mathcal{I}(\theta) = \mathsf{E}_\theta\{\rho_\theta(x)\rho_\theta^*(x)\}$, где $\rho_\theta(x) = \nabla_\theta \ln p_\theta(x)$ — градиент (вектор-столбец) по переменной $\theta$ от функции $\ln p_\theta(x)$.

**Теорема 6.3 (неравенство информации).** *Пусть на $\mathbb{R}^n$ имеется семейство положительных плотностей вероятностей $\{p_\theta(x) \mid \theta \in \mathbb{R}^m\}$ по отношению к некоторой $\sigma$-конечной мере $\mu$, и задана статистическая оценка $\hat\theta = T(X)$ для параметра $\theta$. Если выражения $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx)$ и $\int_{\mathbb{R}^{nN}} T(X) p_\theta(X)\,\mu^N(dX)$ можно дифференцировать по $\theta$ под знаком интеграла, то*

$$\mathsf{E}_\theta\Big\{ \big(\hat\theta - \theta\big)\big(\hat\theta - \theta\big)^* \Big\} \geqslant \big(I + b'(\theta)\big)\big(N\mathcal{I}(\theta)\big)^{-1}\big(I + b'(\theta)\big)^* + b(\theta)b^*(\theta).$$

*Здесь неравенство $A \geqslant B$ для симметричных матриц $A$, $B$ понимается в том смысле, что матрица $A - B$ неотрицательно определена.*

**Задача 2.** Докажите эту теорему по аналогии с теоремой 6.2: начните с аналога леммы 6.1 для матрицы $\mathcal{I}_N(\theta) = \mathsf{E}_\theta\{\rho_\theta(X)\rho_\theta^*(X)\}$, где $\rho_\theta(X) = \nabla_\theta \ln p_\theta(X)$; затем раскройте по линейности неравенство

$$\mathsf{E}_\theta\Big\{ \big(\hat\theta - \mathsf{E}_\theta\hat\theta - C\rho_\theta(X)\big)\big(\hat\theta - \mathsf{E}_\theta\hat\theta - C\rho_\theta(X)\big)^* \Big\} \geqslant 0$$

с произвольной $(m \times m)$-матрицей $C$; а потом положите $C = (I + b'(\theta))\mathcal{I}_N^{-1}(\theta)$.

# § 7. Эффективные оценки

Естественно считать статистическую оценку $\hat\theta$ параметра $\theta$ тем лучше, чем меньше ее вариация $\operatorname{Var}\hat\theta = \mathsf{E}_\theta\big\{(\hat\theta - \theta)^2\big\}$. Однако в силу неравенства Рао — Крамéра вариацию невозможно сделать сколь угодно малой. Например, если оценка несмещенная и выполняются условия теоремы 6.2, то в любом случае $\operatorname{Var}\hat\theta \geqslant 1/\mathcal{I}_N(\theta)$.

**Определение.** Оценка $\hat{\theta} = T(X)$ называется *эффективной,* если она несмещенная и ее вариация имеет минимально возможное значение $\operatorname{Var}\hat{\theta} = 1/\mathcal{I}_N(\theta)$. А если $\operatorname{Var}\hat{\theta} \cdot \mathcal{I}_N(\theta) \to 1$ при $N \to \infty$, то такая оценка называется *асимптотически эффективной* (несмещенность в этом случае не требуется).

В § 2 мы доказали, что если $\operatorname{Var}\hat{\theta}$ стремится к нулю, то оценка $\hat{\theta}$ состоятельна (см. свойство 2.3 вариаций). Отсюда следует, что эффективные и асимптотически эффективные оценки состоятельны.

**Теорема 7.1.** *Если в условиях теоремы 6.2 функция $\mathcal{I}(\theta)$ строго положительна, то эффективность оценки $\hat{\theta} = T(X)$ параметра $\theta$ равносильна тому, что для некоторой функции $d_N(\theta)$*

$$\hat{\theta} - \theta = d_N(\theta)\frac{d\ln p_\theta(X)}{d\theta} \quad \text{п. н.} \tag{7.1}$$

*А если выполняется равенство* (7.1)*, то тогда $d_N(\theta) = 1/\mathcal{I}_N(\theta)$.*

Доказательство. Для несмещенной оценки выкладка (6.6) принимает вид

$$\mathsf{E}_\theta\Big\{\big(\hat{\theta} - \theta - c\rho_\theta(X)\big)^2\Big\} = \mathsf{E}_\theta\Big\{\big(\hat{\theta} - \theta\big)^2\Big\} - 2c + c^2\mathcal{I}_N(\theta) \geqslant 0, \tag{7.2}$$

где $\rho_\theta(X) = d\ln p_\theta(X)/d\theta$. С другой стороны, эффективность несмещенной оценки $\hat{\theta}$ равносильна тому, что неравенство (7.2) обращается в равенство при $c = 1/\mathcal{I}_N(\theta)$. Если это происходит, то $\hat{\theta} - \theta - c\rho_\theta(X) = 0$ с вероятностью 1, что доказывает (7.1).

Наоборот, пусть выполняется (7.1). Напомним, что $\mathsf{E}_\theta\{\rho_\theta(X)\} = 0$. Поэтому оценка $\hat{\theta}$ будет несмещенной. Возьмем $c = d_N(\theta)$. Тогда неравенство (7.2) обратится в равенство. С другой стороны, выражение (7.2) принимает свое минимальное значение при $c = 1/\mathcal{I}_N(\theta)$. Отсюда вытекает, что $d_N(\theta) = 1/\mathcal{I}_N(\theta)$ и оценка $\hat{\theta}$ эффективна. □

Эта теорема может создать неверное впечатление, что эффективная оценка всегда существует и имеет вид $T(X) = \theta + \mathcal{I}_N^{-1}(\theta)d\ln p_\theta(X)/d\theta$. Однако последнее равенство возможно только в том случае, когда его правая часть *не зависит от* $\theta$. Это очень жесткое условие; оно может выполняться лишь для немногих семейств распределений специального вида (см. пример ниже). В общем случае эффективной оценки не существует. Пример неэффективной оценки приведен на с. 30.

Гораздо лучше обстоит дело с асимптотически эффективными оценками. При выполнении некоторых естественных предположений относительно семейства плотностей распределений $p_\theta(x)$ асимптотически эффективными оказываются оценки максимального правдоподобия, которые будут рассмотрены в следующем параграфе.

**Пример.** Рассмотрим серию испытаний Бернулли $X = (x_1, \ldots, x_N)$ с вероятностью успеха $\theta$ (другими словами, каждое $x_t$ принимает значение 1 с вероятностью $\theta$ и значение 0 с вероятностью $1-\theta$). Требуется проверить эффективность оценки $\bar{x}$ для $\theta$.

*Решение.* Для дискретной случайной величины $x_t$ плотность распределения (по отношению к считающей мере, сосредоточенной в двух точках 0 и 1) совпадает с вероятностью. Она может быть записана в виде $p_\theta(x_t) = \theta^{x_t}(1-\theta)^{1-x_t}$. Плотность распределения всей выборки $X$ будет равна

$$p_\theta(X) = \prod_{t=1}^{N} p_\theta(x_t) = \theta^{N\bar{x}}(1-\theta)^{N-N\bar{x}}.$$

Следовательно,

$$\frac{d\ln p_\theta(X)}{d\theta} = \frac{N\bar{x}}{\theta} - \frac{N - N\bar{x}}{1-\theta} = \frac{N}{\theta(1-\theta)}(\bar{x} - \theta).$$

Мы получили равенство (7.1), в котором $\hat{\theta} = \bar{x}$ и $d_N(\theta) = \theta(1-\theta)/N$. Значит, оценка $\bar{x}$ эффективна.

# § 8. Метод максимального правдоподобия

Этот метод построения статистических оценок был предложен в 1912 г. английским исследователем Р. Фишером.

Пусть на пространстве $\mathbb{R}^n$ задана $\sigma$-конечная мера $\mu$ и семейство плотностей вероятности $p_\theta(x)$ по отношению к $\mu$. Рассмотрим выборку $X = (x_1, \ldots, x_N)$ из распределения с плотностью $p_\theta(x)$. *Функцией правдоподобия Фишера* называется плотность вероятности выборки $X$, рассматриваемая как функция от параметра $\theta$:

$$L(\theta) = p_\theta(X) = \prod_{t=1}^{N} p_\theta(x_t). \tag{8.1}$$

*Логарифмической* функцией правдоподобия называется

$$l(\theta) = \ln p_\theta(X) = \sum_{t=1}^{N} \ln p_\theta(x_t). \tag{8.2}$$

Очевидно, в случае считающей меры $\mu$ функция правдоподобия $L(\theta)$ совпадает с вероятностью наблюдения выборки $X$ в зависимости от $\theta$.

*Оценкой максимального правдоподобия* (ОМП) называется такое значение $\hat{\theta} = T(X)$, при котором достигается максимум $L(\theta)$ (или, что то же самое, $l(\theta)$). Если функция $l(\theta)$ дифференцируема, а область изменения параметра $\theta$ открыта, то ОМП должна удовлетворять *уравнению правдоподобия*

$$l'(\theta) = 0. \tag{8.3}$$

Сразу оговоримся, что возможна ситуация, когда в открытой области изменения $\theta$ оценка максимального правдоподобия не существует ни при каком $X$. С другой стороны, если область изменения параметра $\theta$ компактна, а функция $L(\theta)$ непрерывна, то существование ОМП гарантирует теорема Вейерштрасса.

**Пример 1.** Пусть $p(x)$ — некоторая фиксированная всюду положительная плотность вероятности на вещественной прямой. Рассмотрим семейство плотностей вероятности, зависящее от двумерного параметра $\theta = (\mu, \sigma)$:

$$p_\theta(x) = \frac{1}{2\sigma}\, p\Big(\frac{x-\mu}{\sigma}\Big) + \frac{1}{2}\, p(x), \qquad \sigma > 0.$$

Его функция правдоподобия удовлетворяет соотношениям

$$L(\mu, \sigma) = \prod_{t=1}^{N} p_\theta(x_t) > \frac{1}{2\sigma}\, p\Big(\frac{x_1-\mu}{\sigma}\Big) \prod_{t=2}^{N} \frac{1}{2}\, p(x_t).$$

Очевидно, при $\mu = x_1$ и $\sigma \to 0$ выражение в правой части последнего неравенства неограниченно возрастает. Поэтому ОМП не существует.

Р. Фишер доказал, что при некоторых ограничениях на семейство $p_\theta(x)$, включающих компактность области изменения $\theta$, оценка максимального правдоподобия является сильно состоятельной, асимптотически нормальной и асимптотически эффективной. К сожалению, доказательство асимптотической эффективности ОМП довольно сложно (его

можно прочитать, например, в книге Боровкова [1]). Мы здесь докажем только состоятельность и асимптотическую нормальность ОМП.

Предположим, что семейство вероятностных распределений

$$\bigl\{p_\theta(x)\mu(dx)\,\big|\,\theta\in[a,b]\bigr\}$$

удовлетворяет следующим *условиям регулярности*:

$R1$) выражение $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx)$ можно дважды дифференцировать по $\theta$ под знаком интеграла;

$R2$) информационная функция

$$\mathcal{I}(\theta)=\mathsf{E}_\theta\left\{\left(\frac{p'_\theta(x)}{p_\theta(x)}\right)^2\right\}=\mathsf{E}_\theta\left\{\left(\frac{d\ln p_\theta(x)}{d\theta}\right)^2\right\}$$

конечна и строго положительна при всех $\theta\in[a,b]$;

$R3$) существуют такие функция $H(x)$ и константа $M$, что при всех $\theta\in[a,b]$

$$\left|\frac{d^3\ln p_\theta(x)}{d\theta^3}\right|\leqslant H(x),\qquad \mathsf{E}_\theta\{H(x)\}\leqslant M.$$

**Предложение 8.1.** *Если в условиях регулярности $R1$), $R2$) существует эффективная оценка $\hat\theta$ для $\theta$, то она удовлетворяет уравнению правдоподобия.*

Доказательство. В силу теоремы 7.1 для эффективной оценки $\hat\theta$ выполняется равенство $\hat\theta-\theta=d_N(\theta)l'(\theta)$, где $d_N(\theta)=1/\mathcal{I}_N(\theta)$. Из него видно, что $l'(\hat\theta)=0$. $\square$

Это предложение и теорема 7.1 дают способ нахождения эффективных оценок (если они вообще существуют). Вначале следует найти решение уравнения правдоподобия $\hat\theta$. Затем нужно выписать тождество $\hat\theta-\theta=d_N(\theta)l'(\theta)$. Если в нем коэффициент $d_N(\theta)$ не зависит от выборки, то оценка $\hat\theta$ эффективна, а если $d_N(\theta)$ зависит от выборки, то эффективной оценки не существует.

**Пример 2.** Дано семейство нормальных распределений $\mathcal{N}(\mu,\sigma^2)$ с математическим ожиданием $\mu$ и дисперсией $\sigma^2$. Требуется найти эффективные оценки для $\mu$ и для $\sigma^2$ (при оценивании одного параметра другой считается известным).

*Решение.* Логарифмическая функция правдоподобия имеет вид

$$l\big(\mu,\sigma^2\big)=\ln\prod_{t=1}^{N}\frac{e^{-(x_t-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}=-\frac{N}{2}\ln 2\pi-\frac{N}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(x_t-\mu)^2.$$

Составляем и решаем уравнения правдоподобия для $\mu$:

$$\frac{\partial l\big(\mu,\sigma^2\big)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{t=1}^{N}(x_t-\mu)=0\quad\Longrightarrow\quad\hat\mu=\frac{1}{N}\sum_{t=1}^{N}x_t=\bar x;$$

и для $\sigma^2$:

$$\frac{\partial l\big(\mu,\sigma^2\big)}{\partial\sigma^2}=-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{t=1}^{N}(x_t-\mu)^2=0\quad\Longrightarrow\quad\hat\sigma^2=\frac{1}{N}\sum_{t=1}^{N}(x_t-\mu)^2.$$

Выписываем тождество $\hat\theta-\theta=d_{_N}l'(\theta)$ для $\theta=\mu$ и для $\theta=\sigma^2$:

$$\bar x-\mu=d_{_N}\frac{1}{\sigma^2}\sum_{t=1}^{N}(x_t-\mu),$$

$$\frac{1}{N}\sum_{t=1}^{N}(x_t-\mu)^2-\sigma^2=d_{_N}\left(-\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{t=1}^{N}(x_t-\mu)^2\right).$$

Из первого тождества находим $d_{_N}=\sigma^2/N$, а из второго $d_{_N}=2\sigma^4/N$. В обоих случаях множитель $d_{_N}$ не зависит от выборки. Значит, найденные оценки эффективны.

Из этого примера следует, что оценка $\hat\sigma^2=\frac{1}{N}\sum_{t=1}^{N}(x_t-\bar x)^2$ не эффективна (на самом деле она асимптотически эффективна).

**Задача 1.** Докажите, что в рассмотренном примере оценка максимального правдоподобия для *двумерного* параметра $\theta=(\mu,\sigma^2)$ имеет вид $\hat\theta=(\hat\mu,\hat\sigma^2)$, где $\hat\mu=\bar x$ и $\hat\sigma^2=\frac{1}{N}\sum_{t=1}^{N}(x_t-\bar x)^2$.

**Лемма 8.2.** *В условиях регулярности* $R1$), $R2$)

$$\mathsf{E}_\theta\left\{\frac{d^2\ln p_\theta(x)}{d\theta^2}\right\}=-\mathcal{I}(\theta).$$

Доказательство. Дважды дифференцируя по $\theta$ равенство

$$\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx) = 1,$$

получаем

$$\int_{\mathbb{R}^n} p''_\theta(x)\,\mu(dx) = \int_{\mathbb{R}^n} \frac{p''_\theta(x)}{p_\theta(x)} p_\theta(x)\,\mu(dx) = \mathsf{E}_\theta\left\{\frac{p''_\theta(x)}{p_\theta(x)}\right\} = 0.$$

Поэтому

$$\mathsf{E}_\theta\left\{\frac{d^2 \ln p_\theta(x)}{d\theta^2}\right\} = \mathsf{E}_\theta\left\{\frac{d}{d\theta}\frac{p'_\theta(x)}{p_\theta(x)}\right\} =$$

$$= \mathsf{E}_\theta\left\{\frac{p''_\theta(x)}{p_\theta(x)}\right\} - \mathsf{E}_\theta\left\{\left(\frac{p'_\theta(x)}{p_\theta(x)}\right)^2\right\} = -\mathcal{I}(\theta). \quad \square$$

**Теорема 8.3.** *Если выполняются условия регулярности* $R1$), $R2$), $R3$), *и истинное значение параметра* $\theta = \theta_0$ *лежит строго внутри отрезка* $[a, b]$, *то с вероятностью* 1 *при всех достаточно больших* $N$ *уравнение правдоподобия* $l'(\theta) = 0$ *имеет решение* $\hat\theta$, *которое сходится к* $\theta_0$ *и асимптотически нормально распределено*:

$$P_{\theta_0}\left\{\left(\hat\theta - \theta_0\right)\sqrt{N\mathcal{I}(\theta_0)} \leqslant z\right\} \longrightarrow \Phi(z). \tag{8.4}$$

Доказательство. Напомним равенство (8.2)

$$l(\theta) = \sum_{t=1}^{N} \ln p_\theta(x_t).$$

Разложим функцию $l'(\theta)/N$ по Тейлору в окрестности точки $\theta = \theta_0$:

$$\frac{l'(\theta)}{N} = B_1 + B_2(\theta - \theta_0) + B_3\frac{(\theta - \theta_0)^2}{2}, \tag{8.5}$$

где

$$B_1 = \frac{1}{N}\sum_{t=1}^{N} \frac{d \ln p_\theta(x_t)}{d\theta}\bigg|_{\theta=\theta_0}, \qquad B_2 = \frac{1}{N}\sum_{t=1}^{N} \frac{d^2 \ln p_\theta(x_t)}{d\theta^2}\bigg|_{\theta=\theta_0},$$

$$B_3 = \frac{1}{N}\sum_{t=1}^{N} \frac{d^3 \ln p_\theta(x_t)}{d\theta^3}\bigg|_{\theta=\tau}, \qquad \tau \in (\theta_0, \theta).$$

Из леммы 6.1 следует, что $\mathsf{E}B_1 = 0$, а из леммы 8.2 следует, что $\mathsf{E}B_2 = -\mathcal{I}(\theta_0)$. По усиленному закону больших чисел $B_1 \to 0$ почти наверное и $B_2 \to -\mathcal{I}(\theta_0)$ почти наверное. В силу условия регулярности $R3$) величину $B_3$ можно записать в виде

$$B_3 = \frac{\Delta}{N}\sum_{t=1}^{N} H(x_t) = \Delta B_4, \qquad B_4 = \frac{1}{N}\sum_{t=1}^{N} H(x_t),$$

где $|\Delta| \leqslant 1$. По закону больших чисел случайная величина $B_4$ почти наверное сходится к константе $\mathsf{E}\{H(x_t)\}$, которая в силу условия $R3$) не превосходит $M$.

Фиксируем любое положительное число $\varepsilon \leqslant \mathcal{I}(\theta_0)/M$. Тогда почти наверное

$$\begin{aligned}
\overline{\lim_{N\to\infty}}\ \frac{l'(\theta_0+\varepsilon)}{N} &= \overline{\lim_{N\to\infty}}\left\{B_1 + \varepsilon B_2 + \frac{\varepsilon^2}{2}\Delta B_4\right\} \leqslant \\
&\leqslant -\varepsilon\mathcal{I}(\theta_0) + \frac{\varepsilon^2}{2}M \leqslant -\frac{\varepsilon}{2}\mathcal{I}(\theta_0) < 0, \\
\underline{\lim_{N\to\infty}}\ \frac{l'(\theta_0-\varepsilon)}{N} &= \underline{\lim_{N\to\infty}}\left\{B_1 - \varepsilon B_2 + \frac{\varepsilon^2}{2}\Delta B_4\right\} \geqslant \\
&\geqslant \varepsilon\mathcal{I}(\theta_0) - \frac{\varepsilon^2}{2}M \geqslant \frac{\varepsilon}{2}\mathcal{I}(\theta_0) > 0.
\end{aligned}$$

Отсюда вытекает, что почти наверное при всех достаточно больших $N$ функция $l'(\theta)$ имеет разные знаки в точках $\theta_0 \pm \varepsilon$. По теореме о промежуточном значении уравнение правдоподобия $l'(\theta) = 0$ имеет решение $\hat{\theta} \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$. Приравнивая выражение (8.5) к нулю, получаем

$$\hat{\theta} - \theta_0 = \frac{B_1}{-B_2 - \bigl(\hat{\theta} - \theta_0\bigr)\Delta B_4/2} \xrightarrow{\text{п. н.}} 0. \tag{8.6}$$

Значит, оценка $\hat{\theta}$ сильно состоятельна. Далее, из (8.6) следует, что

$$\bigl(\hat{\theta} - \theta_0\bigr)\sqrt{N\mathcal{I}(\theta_0)} = \frac{B_1\sqrt{N/\mathcal{I}(\theta_0)}}{-B_2/\mathcal{I}(\theta_0) - \bigl(\hat{\theta} - \theta_0\bigr)\Delta B_4/2\mathcal{I}(\theta_0)}. \tag{8.7}$$

В силу центральной предельной теоремы распределение числителя в правой части (8.7) сходится к стандартному нормальному распределению $\mathcal{N}(0,1)$, а знаменатель почти наверное сходится к 1. Поэтому распределение всего выражения (8.7) сходится к $\mathcal{N}(0,1)$. $\square$

**Задача 2.** Аккуратно проверьте последнее утверждение.

Только что доказанная теорема требует некоторого обсуждения. Нужно понимать, что из нее *не* вытекает асимптотическая эффективность ОМП (как пишут авторы некоторых учебников, в том числе и сам Гаральд Крамéр в своей книге, изданной в Париже в 1946 году). Дело в том, что из сходимости распределений не вытекает сходимость (и даже существование) дисперсий. Кроме того, сильная состоятельность и асимптотическая нормальность ОМП были доказаны в предположении, что область изменения параметра $\theta$ ограничена интервалом $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, где $\varepsilon \leqslant \mathcal{I}(\theta_0)/M$. Приведем без доказательства усиленный вариант теоремы 8.3.

**Теорема 8.4** [1]. *Если при* $\theta \in [a, b]$ *все плотности* $p_\theta(x)$ *разные, выполняются условия регулярности* $R1$), $R2$), $R3$), *и истинное значение параметра* $\theta$ *лежит в интервале* $(a, b)$, *то оценка максимального правдоподобия для* $\theta$ *на отрезке* $[a, b]$ *сильно состоятельна, асимптотически нормальна и асимптотически эффективна.*

Эта теорема может быть обобщена и на случай многомерного $\theta$.

# § 9. Условные математические ожидания

Для изучения байесовских оценок и достаточных статистик необходимо знакомство с условными математическими ожиданиями и условными распределениями. Вообще говоря, эти темы относятся скорее не к статистике, а к теории вероятностей и теории меры. К сожалению, в университетских курсах теории вероятностей и теории меры их обычно не изучают. Поэтому для полноты изложения мы их обсудим в настоящем и следующем параграфах.

Начнем с определения условных математических ожиданий. Пусть задано некоторое вероятностное пространство $(\Omega, \mathfrak{A}, P)$, и в $\sigma$-алгебре $\mathfrak{A}$ есть $\sigma$-подалгебра $\mathfrak{B}$ (образно говоря, алгебра $\mathfrak{B}$ состоит из более «крупных» множеств, чем $\mathfrak{A}$). Например, если задан случайный вектор $\eta : \Omega \to \mathbb{R}^n$, то в качестве $\mathfrak{B}$ можно взять совокупность прообразов $\eta^{-1}(M)$ всех борелевских множеств $M \subset \mathbb{R}^n$. В этом случае говорят, что $\sigma$-алгебра $\mathfrak{B}$ порождена случайным вектором $\eta$.

Пусть $\xi : \Omega \to \mathbb{R}$ — случайная величина, имеющая конечное математическое ожидание $\mathsf{E}\xi = \int_\Omega \xi \, dP$. Рассмотрим аддитивную функцию

множества $\mu(B) = \int_B \xi\, dP$ на подалгебре $\mathfrak{B}$. Очевидно, она абсолютно непрерывна. По теореме Радона — Никóдима (см. Приложение, п. V) существует такая $\mathfrak{B}$-измеримая функция $\bar{\xi}\colon \Omega \to \mathbb{R}$, для которой

$$\int_B \xi\, dP = \int_B \bar{\xi}\, dP, \qquad B \in \mathfrak{B}. \tag{9.1}$$

Функция $\bar{\xi}$ называется *условным математическим ожиданием* случайной величины $\xi$ по подалгебре $\mathfrak{B}$ и обозначается $\mathsf{E}\{\xi|\mathfrak{B}\}$.

**Пример.** Рассмотрим случай, когда подалгебра $\mathfrak{B} \subset \mathfrak{A}$ порождена разбиением $\Omega$ на конечное или счетное число непересекающихся подмножеств $B_i$. По определению, функция $\bar{\xi} = \mathsf{E}\{\xi|\mathfrak{B}\}$ измерима относительно алгебры $\mathfrak{B}$. Поэтому она постоянна на каждом множестве $B_i$. Тогда из (9.1) следует, что

$$\bar{\xi}(B_i) = \frac{\int_{B_i} \xi\, dP}{P(B_i)}$$

есть среднее интегральное значение функции $\xi$ на $B_i$.

В общем случае тоже полезно представлять себе условное математическое ожидание как результат усреднения функции $\xi$ по элементам подалгебры $\mathfrak{B}$ (и воспринимать (9.1) как строгую формализацию процедуры такого усреднения).

Для изучения свойств условного математического ожидания нам потребуются три леммы.

**Лемма 9.1.** *Если случайная величина $\eta$ измерима относительно $\sigma$-алгебры $\mathfrak{B}$ и для каждого множества $B \in \mathfrak{B}$ выполняется неравенство $\int_B \eta\, dP \geqslant 0$, то почти наверное $\eta \geqslant 0$. А если $\int_B \eta\, dP = 0$ для всех $B \in \mathfrak{B}$, то почти наверное $\eta = 0$.*

**Лемма 9.2.** *Всякая функция $\xi$ представима в виде $\xi = \xi^+ - \xi^-$, где $\xi^\pm = (|\xi| \pm \xi)/2$. При этом $\xi^\pm \geqslant 0$ и $\xi^+ + \xi^- = |\xi|$.*

**Лемма 9.3.** *Каждая неотрицательная $\mathfrak{B}$-измеримая функция $\eta$ может быть представлена как сумма ряда $\eta = \sum_{i=1}^{\infty} c_i \chi_{B_i}$, где $c_i$ — положительные константы, а $\chi_{B_i}$ — характеристические функции некоторых множеств $B_i \in \mathfrak{B}$.*

Доказательство этих лемм обычно приводится в курсах, посвященных интегралу Лебега. Предоставим его читателю.

**Теорема 9.4.** *Условное математическое ожидание* $\bar{\xi} = \mathsf{E}\{\xi|\mathfrak{B}\}$ *обладает следующими свойствами*:

$$\mathsf{E}\xi = \mathsf{E}\bar{\xi}, \tag{9.2}$$

$$\text{если} \quad \xi \geqslant 0 \ \text{ п. н.}, \ \text{то} \quad \bar{\xi} \geqslant 0 \ \text{ п. н.}, \tag{9.3}$$

$$\mathsf{E}|\bar{\xi}| \leqslant \mathsf{E}|\xi|; \tag{9.4}$$

*если случайная величина* $\eta$ *измерима относительно* $\sigma$*-алгебры* $\mathfrak{B}$ *и существуют математические ожидания* $\mathsf{E}\xi$ *и* $\mathsf{E}\{\xi\eta\}$, *то*

$$\mathsf{E}\{\xi\eta\} = \mathsf{E}\{\bar{\xi}\eta\}, \tag{9.5}$$

$$\mathsf{E}\{\xi\eta|\mathfrak{B}\} = \mathsf{E}\{\xi|\mathfrak{B}\}\eta \quad \text{п. н.}; \tag{9.6}$$

*а если к тому же существуют математические ожидания* $\mathsf{E}\{\xi^2\}$ *и* $\mathsf{E}\{\eta^2\}$, *то*

$$\mathsf{E}\{\xi|\mathfrak{B}\}^2 \leqslant \mathsf{E}\{\xi^2|\mathfrak{B}\} \quad \text{п. н.}, \tag{9.7}$$

$$\mathsf{E}\{|\xi - \mathsf{E}\{\xi|\mathfrak{B}\}|^2\} \leqslant \mathsf{E}\{|\xi - \eta|^2\}. \tag{9.8}$$

**Следствие 9.5.** *В случае, когда* $\mathsf{E}\{\xi^2\} < \infty$, *или, иначе говоря,* $\xi \in L^2(\Omega, \mathfrak{A}, P)$, *случайная величина* $\bar{\xi} = \mathsf{E}\{\xi|\mathfrak{B}\}$ *совпадает с ортогональной проекцией* $\xi$ *на подпространство* $L^2(\Omega, \mathfrak{B}, P) \subset L^2(\Omega, \mathfrak{A}, P)$.

Доказательство. Возьмем $B = \Omega$ в тождестве (9.1). Тогда оно превращается в (9.2). Если $\xi \geqslant 0$ п. н., то из (9.1) следует, что $\int_B \bar{\xi}\, dP \geqslant 0$ для любого множества $B \in \mathfrak{B}$. Значит, $\bar{\xi} \geqslant 0$ п. н. (по лемме 9.1). Заметим, что $|\xi| \pm \xi \geqslant 0$. Переходя к условным математическим ожиданиям, получаем неравенства $\mathsf{E}\{|\xi|\,|\mathfrak{B}\} \pm \bar{\xi} \geqslant 0$ п. в. Из них следует, что $|\bar{\xi}| \leqslant \mathsf{E}\{|\xi|\,|\mathfrak{B}\}$, и затем $\mathsf{E}|\bar{\xi}| \leqslant \mathsf{E}\{\mathsf{E}\{|\xi|\,|\mathfrak{B}\}\} = \mathsf{E}|\xi|$.

Равенство (9.5) достаточно доказать для неотрицательных случайных величин $\xi$, $\eta$ (в силу леммы 9.2). Пусть $\xi, \eta \geqslant 0$. По лемме 9.3 представим $\eta$ в виде $\eta = \sum_i c_i \chi_{B_i}$, где $c_i > 0$, а $\chi_{B_i}$ — характеристические функции некоторых множеств $B_i \in \mathfrak{B}$. Тогда с помощью теоремы Леви и (9.1) получаем

$$E\{\xi\eta\} = \int_\Omega \xi \sum_{i=1}^{\infty} c_i \chi_{B_i}\, dP = \sum_{i=1}^{\infty} c_i \int_{B_i} \xi\, dP =$$

$$= \sum_{i=1}^{\infty} c_i \int_{B_i} \bar{\xi}\, dP = \int_\Omega \bar{\xi} \sum_{i=1}^{\infty} c_i \chi_{B_i}\, dP = \mathsf{E}\{\bar{\xi}\eta\}.$$

Заменим в (9.5) функцию $\xi$ на произведение $\xi\eta$, а функцию $\eta$ на $\chi_B$. Тогда

$$\int_B \mathsf{E}\{\xi\eta|\mathfrak{B}\}\,dP = \mathsf{E}\big\{\mathsf{E}\{\xi\eta|\mathfrak{B}\}\chi_B\big\} = \mathsf{E}\{\xi\eta\chi_B\} =$$

$$= \mathsf{E}\big\{\mathsf{E}\{\xi|\mathfrak{B}\}\eta\chi_B\big\} = \int_B \mathsf{E}\{\xi|\mathfrak{B}\}\eta\,dP$$

для всех $B \in \mathfrak{B}$. Отсюда вытекает (9.6).

Далее, в силу (9.3) и (9.6)

$$\mathsf{E}\big\{(\xi-\bar\xi)^2\,\big|\,\mathfrak{B}\big\} = \mathsf{E}\big\{\xi^2 - 2\xi\bar\xi + \bar\xi^2\,\big|\,\mathfrak{B}\big\} = \mathsf{E}\{\xi^2|\mathfrak{B}\} - 2\bar\xi^2 + \bar\xi^2 \geqslant 0$$

почти наверное. Тем самым доказано (9.7).

Из тождества (9.5) следует, что $\mathsf{E}\{(\xi-\bar\xi)\eta\} = 0$ для любых функций $\xi \in L^2(\Omega,\mathfrak{A},P)$ и $\eta \in L^2(\Omega,\mathfrak{B},P)$. Следовательно, $\bar\xi = \mathsf{E}\{\xi|\mathfrak{B}\}$ является ортогональной проекцией $\xi$ на линейное подпространство $L^2(\Omega,\mathfrak{B},P) \subset L^2(\Omega,\mathfrak{A},P)$. А неравенство (9.8) просто означает, что $\bar\xi$ — это ближайшая к $\xi$ точка из $L^2(\Omega,\mathfrak{B},P)$. $\square$

**Замечание.** Для случайных величин $\xi \in L^2(\Omega,\mathfrak{A},P)$ условное математическое ожидание $\bar\xi$ можно определять как ортогональную проекцию $\xi$ на подпространство $L^2(\Omega,\mathfrak{B},P)$. Из этого определения можно вывести все свойства условных математических ожиданий без использования теоремы Радона — Никóдима.

**Задача 1.** Докажите свойства (9.5) – (9.8) для случайных *векторов* $\xi$, $\eta$ одинаковой размерности (понимая $\xi\eta$ как скалярное произведение).

**Задача 2.** Докажите обобщение свойства (9.7), называемое *неравенством Йенсена*: для любой выпуклой функции $U\colon \mathbb{R} \to \mathbb{R}$ и любой функции $\xi \in L^1(\Omega,\mathfrak{A},P)$

$$U\big(\mathsf{E}\{\xi|\mathfrak{B}\}\big) \leqslant \mathsf{E}\big\{U(\xi)\,\big|\,\mathfrak{B}\big\} \quad \text{п. н.}$$

В ситуации, когда $\sigma$-подалгебра $\mathfrak{B} \subset \mathfrak{A}$ порождена случайным вектором $\eta\colon \Omega \to \mathbb{R}^n$, функцию $\bar\xi = \mathsf{E}\{\xi|\mathfrak{B}\}$ принято называть *условным математическим ожиданием* $\xi$ *при фиксированном* $\eta$ и обозначать как $\mathsf{E}\{\xi|\eta\}$.

**Теорема 9.6.** *Если функция* $\xi : \Omega \to \mathbb{R}$ *измерима относительно* $\sigma$*-алгебры на* $\Omega$*, порожденной отображением* $\eta\colon \Omega \to \mathbb{R}^n$*, то найдется такая борелевская функция* $f\colon \mathbb{R}^n \to \mathbb{R}$*, что* $\xi = f \circ \eta$*.*

**Следствие 9.7.** *Для любой случайной величины* $\xi \in L^1(\Omega, \mathfrak{A}, P)$ *и случайного вектора* $\eta\colon \Omega \to \mathbb{R}^n$ *найдется такая борелевская функция* $f\colon \mathbb{R}^n \to \mathbb{R}$*, что* $\mathsf{E}\{\xi|\eta\} = f(\eta)$*.*

Доказательство. По условию для каждого рационального числа $q$ существует такое борелевское множество $B_q \subset \mathbb{R}^n$, что

$$\{\,\omega \in \Omega \mid \xi(\omega) \leqslant q\,\} = \{\,\omega \in \Omega \mid \eta(\omega) \in B_q\,\}.$$

Рассмотрим функцию $f(x) = \inf\{\,q \in \mathbb{Q} \mid x \in B_q\,\}$. Тогда для любого рационального $q$

$$\xi(\omega) \leqslant q \quad \Leftrightarrow \quad \eta(\omega) \in B_q \quad \Leftrightarrow \quad f(\eta(\omega)) \leqslant q. \tag{9.9}$$

Если $\xi(\omega) \neq f(\eta(\omega))$, то найдется рациональное число $q$, при котором $\xi(\omega) < q < f(\eta(\omega))$ или $\xi(\omega) > q > f(\eta(\omega))$, что противоречит (9.9). Следовательно, $\xi(\omega) \equiv f(\eta(\omega))$.

Для любого $c \in \mathbb{R}$ множество $\{\,x \mid f(x) < c\,\} = \bigcup_{q<c} B_q$ борелевское, что доказывает измеримость $f$. $\square$

# § 10. Условные распределения

Вначале введем понятие условного распределения в простейшем случае. Пусть на конечномерном пространстве $\mathbb{R}^{n+m}(z)$ задана плотность вероятности $p(z)$ по отношению к мере Лебега. Предположим, что все координаты $z$ разбиты на две группы: $z = (x, y)$, где

$$x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \qquad y = (y_1, \ldots, y_m) \in \mathbb{R}^m.$$

Тогда плотность $p(z) = p(x, y)$ можно представить в виде произведения

$$p(x, y) = p(x)p(y|x), \tag{10.1}$$

где

$$p(x) = \int_{\mathbb{R}^m} p(x, y)\, dy, \qquad p(y|x) = \frac{p(x, y)}{p(x)}. \tag{10.2}$$

Заметим, что $\int_{\mathbb{R}^n} p(x)\,dx = 1$. Это означает, что $p(x)$ является плотностью вероятности на $\mathbb{R}^n(x)$. Аналогично, если $p(x) \neq 0$, то $p(y|x)$ является плотностью вероятности на $\mathbb{R}^m(y)$ (в случае $p(x) = 0$ определим плотность $p(y|x)$ произвольно). Условным распределением случайной величины $y$ при фиксированном $x$ называется распределение на пространстве $\mathbb{R}^m(y)$, задаваемое плотностью $p(y|x)$.

Условное распределение позволяет сводить кратное интегрирование к повторному: для любой суммируемой функции $f(x, y)$ и любого борелевского множества $M \subset \mathbb{R}^n$

$$\int_{M\times\mathbb{R}^m} f(x,y)p(x,y)\,dxdy = \int_M \left\{\int_{\mathbb{R}^m} f(x,y)p(y|x)\,dy\right\} p(x)\,dx. \quad (10.3)$$

Другое важнейшее свойство условного распределения — это то, что интегрирование по нему дает условное математическое ожидание при фиксированном $x$.

**Теорема 10.1.** *Если $p(y|x)$ — плотность условного распределения $y$ по $x$, то*

$$\int_{\mathbb{R}^m} f(x,y)p(y|x)\,dy = \mathsf{E}\{f\,|\,x\}. \quad (10.4)$$

Доказательство. Из следствия 9.7 и определения условного математического ожидания вытекает, что $\mathsf{E}\{f\,|\,x\}$ — это такая борелевская функция от $x$, что для любого борелевского множества $M \subset \mathbb{R}^n(x)$ выполняется равенство

$$\int_{M\times\mathbb{R}^m} f(x,y)p(x,y)\,dxdy = \int_{M\times\mathbb{R}^m} \mathsf{E}\{f\,|\,x\}p(x,y)\,dxdy. \quad (10.5)$$

Очевидно, интеграл в левой части (10.4) зависит только от $x$. Обозначим его $\bar{f}(x)$. Тогда в силу (10.3)

$$\int_{M\times\mathbb{R}^m} f(x,y)p(x,y)\,dxdy = \int_M \bar{f}(x)p(x)\,dx =$$

$$= \int_M \bar{f}(x)\left\{\int_{\mathbb{R}^m} p(x,y)\,dy\right\} dx = \int_{M\times\mathbb{R}^m} \bar{f}(x)p(x,y)\,dxdy. \quad (10.6)$$

Сравнивая (10.6) и (10.5), мы видим, что $\bar{f}(x) = \mathsf{E}\{f\,|\,x\}$. $\square$

Формулы (10.1) и (10.2) можно записать в симметричной форме

$$p(x,y) = p(y)p(x\,|\,y), \tag{10.7}$$

где

$$p(y) = \int_{\mathbb{R}^n} p(x,y)\,dx, \qquad p(x\,|\,y) = \frac{p(x,y)}{p(y)}. \tag{10.8}$$

Подставляя (10.1) в (10.8), получаем *формулу Байеса* для условных плотностей

$$p(x\,|\,y) = \frac{p(x)p(y\,|\,x)}{\int_{\mathbb{R}^n} p(x)p(y\,|\,x)\,dx}. \tag{10.9}$$

В общем случае, когда у случайных величин $x \in \mathbb{R}^n$ и $y \in \mathbb{R}^m$ нет плотности совместного распределения, условным распределением $y$ при фиксированном $x$ называется такое семейство борелевских вероятностных мер $P(dy\,|\,x)$ на $\mathbb{R}^m(y)$, зависящее от $x$, что для всякой суммируемой функции $f(x,y)$ выполняется тождество

$$\int_{\mathbb{R}^m} f(x,y)\,P(dy\,|\,x) = \mathsf{E}\{f\,|\,x\}. \tag{10.10}$$

К сожалению, доказать существование условного распределения в ситуации, когда у совместного распределения $x$, $y$ нет плотности, не так просто. Его естественно попытаться строить с помощью условных вероятностей. *Условной вероятностью* борелевского множества $M$ из $\mathbb{R}^m(y)$ при фиксированном значении $x$ называется величина

$$P(M\,|\,x) = \mathsf{E}\{\chi_M\,|\,x\}, \tag{10.11}$$

где $\chi_M$ — характеристическая функция множества $M$. Заметим, что если существует условное распределение $P(dy\,|\,x)$, то

$$\int_M P(dy\,|\,x) = \int_{\mathbb{R}^m} \chi_M\,P(dy\,|\,x) = \mathsf{E}\{\chi_M\,|\,x\}. \tag{10.12}$$

Сравнивая (10.12) с (10.11), мы видим, что вероятность множества $M$ по отношению к распределению $P(dy\,|\,x)$ совпадает с условной вероятностью $M$ при данном $x$.

Если случайная величина $x$ принимает конечное или счетное число значений, то можно просто считать, что условное распределение сов-

падает с условной вероятностью, и никаких осложнений при этом не возникает.

Однако в общей ситуации такой наивный подход наталкивается на серьезные трудности. Дело в том, что условное математическое ожидание в правой части (10.11) определено лишь с точностью до множества меры нуль. Это приводит к тому, что для непересекающихся множеств $M_1, M_2 \subset \mathbb{R}^m$ мы можем гарантировать равенство

$$P(M_1 \cup M_2 \,|\, x) = P(M_1 \,|\, x) + P(M_2 \,|\, x)$$

не при всех $x$, а лишь при почти всех. Множество тех $x$, при которых оно нарушается, может быть разным при разных $M_1$ и $M_2$. И может оказаться так, что не будет ни одного значения $x$, при котором оно выполняется для всех пар непересекающихся множеств $M_1$ и $M_2$.

Несмотря на указанное препятствие, можно доказать, что в случае *конечномерных* случайных величин $x$, $y$ условное распределение $y$ по $x$ всегда существует.

**Задача.** Проверьте, что если $P\{x = x_0\} > 0$, то

$$P(M \,|\, x_0) = \frac{P\{y \in M,\, x = x_0\}}{P\{x = x_0\}}.$$

**Теорема 10.2.** *Если случайные величины $x \in \mathbb{R}^n$ и $y \in \mathbb{R}^m$ независимы, то условное распределение $P(dy\,|\,x)$ совпадает с безусловным распределением $P(dy)$. Наоборот, если условное распределение $P(dy\,|\,x)$ не зависит от $x$, то случайная величина $y$ не зависит от $x$, а ее условное распределение совпадает с безусловным.*

Доказательство. Пусть $x$ и $y$ независимы. Для любой интегрируемой функции $f(x, y)$ положим

$$\bar{f}(x) = \int_{\mathbb{R}^m} f(x, y)\, P(dy). \tag{10.13}$$

Тогда для каждого борелевского множества $M \subset \mathbb{R}^n$ будет

$$\int_{M \times \mathbb{R}^m} f(x, y)\, P(dx) P(dy) = \int_M P(dx) \int_{\mathbb{R}^m} f(x, y)\, P(dy) =$$

$$= \int_M \bar{f}(x)\, P(dx) = \int_{M \times \mathbb{R}^m} \bar{f}(x)\, P(dx) P(dy).$$

По определению, это означает, что $\bar{f}(x) = \mathsf{E}\{f\,|\,x\}$. А из (10.13) следует, что $P(dy) = P(dy\,|\,x)$.

Наоборот, пусть условное распределение $P(dy\,|\,x)$ не зависит от $x$. Тогда для любых борелевских множеств $A \subset \mathbb{R}^n$ и $B \subset \mathbb{R}^m$ имеем

$$P\{x \in A,\, y \in B\} = \int_{\mathbb{R}^n} P(dx) \int_{\mathbb{R}^m} \chi_A(x)\chi_B(y)\, P(dy\,|\,x) = P(A)P(B\,|\,x).$$

При $A = \mathbb{R}^n$ отсюда следует, что $P\{y \in B\} = P(B\,|\,x)$. А при произвольном $A$ получаем определение независимости

$$P\{x \in A,\, y \in B\} = P(A)P(B). \quad \square$$

В заключение приведем наиболее общее определение условного распределения. Пусть задано вероятностное пространство $(\Omega, \mathfrak{A}, P)$ и $\sigma$-подалгебра $\mathfrak{B} \subset \mathfrak{A}$. Условным распределением вероятности $P$ по подалгебре $\mathfrak{B}$ называется такое семейство распределений вероятности $P(\,\cdot\,|\,\omega)$ на пространстве $(\Omega, \mathfrak{A})$, зависящее от параметра $\omega \in \Omega$, что для всякой функции $f \in L^1(\Omega, \mathfrak{A}, P)$ выполняется тождество

$$\int_\Omega f(\omega')\, P(d\omega'\,|\,\omega) \equiv \mathsf{E}\{f\,|\,\mathfrak{B}\}.$$

Можно доказать, что если $\Omega$ — полное сепарабельное метрическое пространство с борелевской $\sigma$-алгеброй $\mathfrak{A}$ (например, $\Omega = \mathbb{R}^n$), то условное распределение существует. А в общей ситуации его может и не быть.

# § 11. Байесовские оценки

Пусть на пространстве $\mathbb{R}^n$ задано семейство плотностей вероятностей $\{p(x\,|\,\theta) \mid \theta \in Q\}$, где $Q$ — открытое подмножество в $\mathbb{R}^m$. Пусть на области $Q$ задана другая плотность вероятности $p(\theta)$, которую мы далее будем называть *априорной*. Будем предполагать, что параметр $\theta$ принимает случайные значения в $Q$ с вероятностями, определяемыми априорной плотностью. Задача байесовского оценивания состоит в том, чтобы строить оценки для $\theta$, оптимальные с точки зрения априорного распределения.

Легко видеть, что две плотности $p(\theta)$ и $p(x\,|\,\theta)$ определяют плотность совместного распределения пары $(\theta, x) \in Q \times \mathbb{R}^n$ по формуле $p(\theta, x) = p(\theta)p(x\,|\,\theta)$. При этом $p(x\,|\,\theta)$ является плотностью условного распределения переменной $x$ при фиксированном $\theta$. Плотность распределения выборки $X = (x_1, \ldots, x_N)$ при фиксированном $\theta$ имеет вид

$p(X|\theta) = p(x_1|\theta) \cdot \ldots \cdot p(x_N|\theta)$. Если выборка $X$ задана, то для нее можно вычислить плотность условного распределения параметра $\theta$ по формуле Байеса:

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{\int_Q p(\theta)p(X|\theta)\,d\theta}. \tag{11.1}$$

Распределение в $Q$ с такой плотностью называется *апостериорным*.

Предположим, что статистик, дающий оценку $\hat{\theta}$ для истинного значения $\theta$, уплачивает штраф в размере $w(\theta, \hat{\theta})$. В этом случае *функционалом риска* для оценки $\hat{\theta} = T(X)$ называется полное математическое ожидание штрафа по отношению к совместному распределению пары $(\theta, X)$. Оно вычисляется по формуле

$$r(T) = \int_Q \int_{\mathbb{R}^{nN}} w(\theta, T(X))p(\theta)p(X|\theta)\,d\theta\,dX = \mathsf{E}\big\{w(\theta, T(X))\big\}.$$

Смысл функционала риска состоит в том, что статистик, использующий оценку $T(X)$, будет в среднем уплачивать штраф $r(T)$.

Принцип оптимальности Байеса заключается в том, что в качестве оценки для $\theta$ следует выбирать такую статистику $\hat{\theta} = T(X)$, при которой функционал риска будет минимальным. Такая оценка называется *байесовской*.

Легче всего байесовская оценка находится для квадратичной функции потерь — штрафа вида $w(\theta, \hat{\theta}) = |\hat{\theta} - \theta|^2$.

**Теорема 11.1.** *Для квадратичной функции потерь оценка Байеса* $\hat{\theta}$ *совпадает с условным математическим ожиданием параметра* $\theta$ *при фиксированном значении выборки* $X$ (*апостериорным средним*)

$$\mathsf{E}\{\theta|X\} = \int_Q \theta p(\theta|X)\,d\theta, \tag{11.2}$$

*где* $p(\theta|X)$ — *апостериорная плотность* (11.1).

Доказательство. Для квадратичной функции потерь функционал риска имеет вид $r(T) = \mathsf{E}\big\{|T(X) - \theta|^2\big\}$, где математическое ожидание вычисляется по отношению к совместному распределению $\theta$ и $X$. В теореме 9.4 нами было получено следующее неравенство (9.8) для условных математических ожиданий:

$$\mathsf{E}\big\{|\xi - \mathsf{E}\{\xi|\mathfrak{B}\}|^2\big\} \leqslant \mathsf{E}\big\{|\xi - \eta|^2\big\}.$$

Здесь $\xi$ — произвольная случайная величина, а случайная величина $\eta$ измерима относительно $\sigma$-алгебры $\mathfrak{B}$. Положим $\xi = \theta,\ \eta = T(X)$, а в качестве $\mathfrak{B}$ возьмем $\sigma$-алгебру, порожденную случайным вектором $X$. В результате у нас получится неравенство

$$\mathsf{E}\big\{|\theta - \mathsf{E}\{\theta|X\}|^2\big\} \leqslant \mathsf{E}\big\{|\theta - T(X)|^2\big\}.$$

Поэтому функционал риска $r(T)$ принимает минимальное значение на функции $T(X) = \mathsf{E}\{\theta|X\}$, и она является байесовской оценкой. А формула (11.2) для условного математического ожидания была доказана в теореме 10.1. $\square$

**Задача 1.** Проверьте эти рассуждения для многомерных $\theta$ и $T(X)$.

**Задача 2.** Докажите неравенство $r(T) \geqslant r\big(\mathsf{E}\{\theta|X\}\big)$ посредством прямого вычисления.

**Пример.** Пусть $X = (x_1, \ldots, x_N)$ — выборка из экспоненциального распределения с плотностью $p(x|\lambda) = \lambda e^{-\lambda x}\ \ (x > 0)$, а параметр $\lambda > 0$ имеет априорное распределение с плотностью $p(\lambda) = e^{-\lambda}$. Требуется найти байесовскую оценку $\hat{\lambda}$ для квадратичной функции потерь.

*Решение.* В этом примере

$$p(\lambda)p(X|\lambda) = e^{-\lambda}\prod_{t=1}^{N}\lambda e^{-\lambda x_t} = \lambda^N e^{-(N\bar{x}+1)\lambda}.$$

В соответствии с теоремой 11.1 и формулой (11.1)

$$\hat{\lambda} = \mathsf{E}\{\lambda|X\} = \frac{\int_0^\infty \lambda p(\lambda)p(X|\lambda)\,d\lambda}{\int_0^\infty p(\lambda)p(X|\lambda)\,d\lambda} = \frac{\int_0^\infty \lambda^{N+1}e^{-(N\bar{x}+1)\lambda}\,d\lambda}{\int_0^\infty \lambda^N e^{-(N\bar{x}+1)\lambda}\,d\lambda}.$$

Сделаем в интегралах замену $u = (N\bar{x}+1)\lambda$. Тогда

$$\hat{\lambda} = \frac{1}{N\bar{x}+1}\,\frac{\int_0^\infty u^{N+1}e^{-u}\,du}{\int_0^\infty u^N e^{-u}\,du} = \frac{1}{N\bar{x}+1}\,\frac{\Gamma(N+2)}{\Gamma(N+1)} = \frac{N+1}{N\bar{x}+1}.$$

Интересно, что байесовская оценка $\hat{\lambda}$ всегда будет меньше $N+1$ (потому что $\bar{x} > 0$), в то время как сам параметр $\lambda$ может быть сколь угодно большим.

Если параметр $\theta$ может принимать дискретный набор значений $\theta_1$, $\theta_2$, $\ldots$, $\theta_k$, то часто используют так называемую (0–1)-функцию потерь, равную нулю когда оценка $\hat{\theta}$ совпадает с истинным значением $\theta_i$ и единице в прочих случаях. Функционал риска в такой ситуации имеет смысл вероятности ошибки при определении истинного значения параметра. В § 20 мы докажем, что соответствующая байесовская оценка будет равна тому $\theta_i$, которому отвечает максимальная апостериорная вероятность $P(\theta_i|X)$.

## § 12. Достаточные статистики†

Понятие достаточной статистики было введено Фишером в 1922 г. Пусть на пространстве $\mathbb{R}^n$ задано семейство распределений вероятности $\{P_\theta \mid \theta \in Q\}$, и пусть $X = (x_1, \ldots, x_N) \in \mathbb{R}^{nN}$ — выборка из распределения $P_\theta$.

**Определение.** Борелевская функция $T : \mathbb{R}^{nN} \to \mathbb{R}^k$ называется *достаточной статистикой* для семейства распределений $P_\theta$ в том случае, когда условное распределение выборки $X$ при фиксированном значении $T(X)$ не зависит от параметра $\theta$.

**Задача 1*.** Докажите, что статистика $T(X)$ является достаточной тогда и только тогда, когда для любого априорного распределения вероятностей $P(d\theta)$ на множестве параметров апостериорное распределение $P(d\theta|X)$ зависит только от $T(X)$.

Содержательный смысл достаточной статистики состоит в том, что если нам стало известно значение $T(X)$, то из выборки $X$ уже невозможно извлечь никакой дополнительной информации о параметре $\theta$. Это неформальное утверждение иллюстрирует следующая теорема.

**Теорема 12.1 (Колмогоров — Блэкуэлл — Рао).** *Если $T(X)$ — достаточная статистика для некоторого семейства распределений вероятности $P_\theta$, то для любой функции $f(\theta)$, любой ее оценки $\hat{f}(X)$ и функции $\bar{f}(T(X)) = \mathsf{E}_\theta\{\hat{f}(X)\,|\,T(X)\}$ имеет место неравенство*

$$\mathsf{E}_\theta\Big\{\big(\bar{f}(T(X)) - f(\theta)\big)^2\Big\} \leqslant \mathsf{E}_\theta\Big\{\big(\hat{f}(X) - f(\theta)\big)^2\Big\}.$$

Из этой теоремы следует, что для построения статистической оценки с минимальной вариацией достаточно знать не всю выборку $X$, а лишь значение $T(X)$.

Доказательство. По свойству (9.7) условных математических ожиданий

$$\bigl(\bar f(T(X)) - f(\theta)\bigr)^2 = \bigl(\mathsf{E}_\theta\bigl\{\hat f(X)\,\big|\,T(X)\bigr\} - f(\theta)\bigr)^2 =$$

$$= \mathsf{E}_\theta\bigl\{\hat f(X) - f(\theta)\,\big|\,T(X)\bigr\}^2 \leqslant \mathsf{E}_\theta\Bigl\{\bigl(\hat f(X) - f(\theta)\bigr)^2\,\Big|\,T(X)\Bigr\} \quad \text{п. н.}$$

Интегрируя левую и правую части этого неравенства, получаем

$$\mathsf{E}_\theta\Bigl\{\bigl(\bar f(T(X)) - f(\theta)\bigr)^2\Bigr\} \leqslant \mathsf{E}_\theta\Bigl\{\mathsf{E}_\theta\Bigl\{\bigl(\hat f(X) - f(\theta)\bigr)^2\,\Big|\,T(X)\Bigr\}\Bigr\} =$$

$$= \mathsf{E}_\theta\Bigl\{\bigl(\hat f(X) - f(\theta)\bigr)^2\Bigr\}. \quad \square$$

**Вопрос.** Где в этом доказательстве использовалась достаточность статистики $T(X)$?

**Задача 2.** Докажите неравенство

$$\mathsf{E}_\theta\bigl\{U\bigl(\bar f(T(X)) - f(\theta)\bigr)\bigr\} \leqslant \mathsf{E}_\theta\bigl\{U\bigl(\hat f(X) - f(\theta)\bigr)\bigr\}$$

для любой выпуклой функции $U$.

Для нахождения достаточных статистик очень полезен следующий критерий. Предположим, что на пространстве $\mathbb{R}^n$ задана какая-нибудь $\sigma$-конечная мера $\mu$ (например, мера Лебега или считающая мера) и семейство плотностей вероятностей $\{p_\theta(x) \mid \theta \in Q\}$ по отношению к ней. Тогда случайная выборка $X = (x_1, \ldots, x_N)$ будет иметь плотность распределения $p_\theta(X) = p_\theta(x_1)\ldots p_\theta(x_N)$ по отношению к мере $\mu^N(dX) = \mu(dx_1)\ldots\mu(dx_N)$.

**Теорема 12.2 (критерий факторизации Неймана – Фишера).** *Статистика $T(X) \in \mathbb{R}^k$ достаточна тогда и только тогда, когда при почти всех $X$ выполняется равенство $p_\theta(X) = g(T(X), \theta)h(X)$, где $g(t, \theta)$ и $h(X)$ — некоторые неотрицательные функции, измеримые соответственно по переменным $t \in \mathbb{R}^k$ и $X \in \mathbb{R}^{nN}$.*

Доказательство. Предположим вначале, что $\mu$ — считающая мера. В этом случае плотность вероятности $p_\theta(X)$ совпадает с вероятностью $P_\theta(X)$. Если статистика $T(X)$ достаточна, то положим

$$g(t, \theta) = P_\theta\{T(X) = t\}, \qquad h(X_0) = P_\theta(X_0 | T(X) = T(X_0)).$$

Фиксируем произвольную выборку $X_0$. Для нее по формуле полной вероятности

$$P_\theta(X_0) = P_\theta\{T(X) = T(X_0)\}P_\theta\big(X_0 \,\big|\, T(X) = T(X_0)\big) = \\ = g\big(T(X_0), \theta\big)h(X_0).$$

Наоборот, пусть $P_\theta(X) = g(T(X), \theta)h(X)$. Тогда

$$P_\theta\big(X_0 \,\big|\, T(X_0)\big) = \frac{P_\theta(X_0)}{P_\theta\{T(X) = T(X_0)\}} = \frac{P_\theta(X_0)}{\sum_{T(X)=T(X_0)} P_\theta(X)} = \\ = \frac{g(T(X_0), \theta)h(X_0)}{\sum_{T(X)=T(X_0)} g(T(X), \theta)h(X)} = \frac{h(X_0)}{\sum_{T(X)=T(X_0)} h(X)},$$

откуда видно, что $P_\theta(X_0\,|\,T(X_0))$ не зависит от $\theta$.

Предположим теперь, что $\mu$ — это мера Лебега, плотность $p_\theta(X)$ всюду положительна, а рассматриваемая статистика

$$T(X) = \big(T_1(X), \ldots, T_k(X)\big)$$

непрерывно дифференцируема и может быть дополнена до криволинейной системы координат на $\mathbb{R}^{nN}$. Это означает, что существует такой набор функций $S(X) = (S_1(X), \ldots, S_{nN-k}(X))$, что отображение $X \mapsto (S(X), T(X))$ из $\mathbb{R}^{nN}$ в $\mathbb{R}^{nN}$ непрерывно дифференцируемо, взаимно однозначно, и обратное к нему отображение $X = X(S, T)$ тоже непрерывно дифференцируемо. Обозначим через $J(S, T)$ якобиан отображения $X(S, T)$. Тогда плотность вероятности в координатах $(S, T)$ имеет вид

$$q_\theta(S, T) = p_\theta(X(S, T))\,|J(S, T)|\,. \tag{12.1}$$

Если $p_\theta(X) = g(T(X), \theta)h(X)$, то

$$q_\theta(S, T) = g(T, \theta)h(X(S, T))\,|J(S, T)|\,.$$

Отсюда следует, что плотность условного распределения

$$q_\theta(S\,|\,T) = \frac{q_\theta(S, T)}{\int_{\mathbb{R}^{nN-k}} q_\theta(S, T)\,dS} = \frac{h(X(S, T))\,|J(S, T)|}{\int_{\mathbb{R}^{nN-k}} h(X(S, T))\,|J(S, T)|\,dS}$$

не зависит от $\theta$. Наоборот, если условная плотность $q_\theta(S|T)$ не зависит от параметра $\theta$, то плотность $p_\theta(X)$ в силу (12.1) и (10.1) представляется в виде

$$p_\theta(X) = \frac{q_\theta(S,T)}{|J(S,T)|} = \frac{q_\theta(T)q_\theta(S|T)}{|J(S,T)|} = g(T(X),\theta)h(X),$$

где $g(T,\theta) = q_\theta(T)$ и $h(X) = q_\theta(S|T)\big/\,|J(S,T)|$.

Полное доказательство этой теоремы под силу разве что занудам-отличникам. Оно изложено в следующем параграфе. $\square$

Довольно часто в приложениях встречаются семейства плотностей вероятностей (по отношению к мере Лебега или считающей мере), которые можно записать в виде

$$p_\theta(x) = \exp\left\{\sum_{j=1}^{k} a_j(\theta)w_j(x) + b(\theta)\right\}h(x).$$

Их называют *экспоненциальными семействами*. Так выглядят, например, семейства нормальных, экспоненциальных, геометрических, биномиальных распределений, распределений Бернулли и Пуассона. Для экспоненциальных семейств достаточные статистики строятся очень легко. Поскольку

$$p_\theta(X) = \prod_{t=1}^{N} p_\theta(x_t) = \exp\left\{\sum_{j=1}^{k} a_j(\theta)\sum_{t=1}^{N} w_j(x_t) + Nb(\theta)\right\}\prod_{t=1}^{N} h(x_t),$$

достаточной будет статистика $T(X) = (T_1(X),\ldots,T_k(X))$, в которой $T_j(X) = \frac{1}{N}\sum_{t=1}^{N} w_j(x_t)$.

**Пример 1.** Найти достаточную статистику для семейства нормальных распределений $\mathcal{N}(a,\sigma^2)$, зависящего от параметра $\theta = (a,\sigma^2)$.

*Решение.* Выпишем плотность распределения выборки $X$:

$$p_\theta(X) = \prod_{t=1}^{N}\frac{e^{-(x_t-a)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(x_t-a)^2\right\} =$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{N}{2\sigma^2}\left(\frac{1}{N}\sum_{t=1}^{N}x_t^2 - \frac{2a}{N}\sum_{t=1}^{N}x_t + a^2\right)\right\}.$$

Очевидно, $p_\theta(X)$ является функцией от переменных $\bar{x}$, $\hat{\sigma}^2$, $a$, $\sigma^2$, где

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N} x_t^2 - \bar{x}^2 = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})^2.$$

По критерию факторизации статистика $T(X) = (\bar{x}, \hat{\sigma}^2)$ достаточна.

**Пример 2.** Для семейства плотностей $p_\theta(x) = e^{\theta - x}$ при $x \geqslant \theta$ и $p_\theta(x) = 0$ при $x < \theta$ найти одномерную достаточную статистику (эти плотности задают сдвиги экспоненциального распределения с показателем $\lambda = 1$).

*Решение.* Представим плотность $p_\theta(x)$ в виде $p_\theta(x) = e^{\theta - x}\eta(x - \theta)$, где $\eta(x)$ — функция Хевисайда. Тогда

$$p_\theta(X) = \prod_{t=1}^{N} e^{\theta - x_t}\eta(x_t - \theta) = e^{N\theta}\eta\Big(\min_t x_t - \theta\Big)\prod_{t=1}^{N} e^{-x_t}.$$

Определим функции $T(X) = \min_t x_t$,

$$g(T(X), \theta) = e^{N\theta}\eta(T(X) - \theta), \qquad h(X) = \prod_{t=1}^{N} e^{-x_t}.$$

Для них выполняется равенство $p_\theta(X) = g(T(X), \theta)h(X)$. Значит, статистика $T(X)$ достаточна. Этот пример опровергает «теорему Дармуа», сформулированную на с. 64 учебника [3] (в которой утверждается, что всякое семейство плотностей, допускающее достаточную статистику той же размерности, что и параметр $\theta$, экспоненциально).

**Задача 3.** Проверьте, что семейства нормальных, геометрических, биномиальных распределений, распределений Бернулли и Пуассона — экспоненциальные (эти распределения описаны в Приложении, п. VI).

**Задача 4.** Докажите, что оценка максимального правдоподобия для параметра $\theta$ зависит лишь достаточной статистики $T(X)$.

**Задача 5.** Докажите, что если существует эффективная оценка для параметра $\theta$, то она является достаточной статистикой.

# § 13. Доказательство критерия факторизации†

**Лемма 13.1.** *Пусть на $\mathbb{R}^n$ задана $\sigma$-конечная борелевская мера $\mu$ и семейство плотностей вероятности $\{p_\theta(x) \mid \theta \in Q\}$ по отношению к ней. Тогда существуют такое борелевское множество $M_0 \subset \mathbb{R}^n$ и счетный набор параметров $\theta_i \in Q$, что а) функция $\sup_i p_{\theta_i}(x)$ строго положительна на $M_0$ и б) для любого $\theta \in Q$ плотность $p_\theta(x)$ обращается в нуль почти всюду на $\mathbb{R}^m \setminus M_0$ (в смысле меры $\mu$).*

Доказательство. Без ограничения общности можно считать, что мера $\mu$ конечна. Отнесем борелевское множество $M \subset \mathbb{R}^n$ к классу $\mathcal{A}$, если существует такой счетный набор параметров $\theta_i \in Q$, что функция $\sup_i p_{\theta_i}(x)$ строго положительна на $M$. Очевидно, объединение счетного числа множеств из $\mathcal{A}$ тоже принадлежит $\mathcal{A}$.

Выберем такую последовательность множеств $M_i \in \mathcal{A}$, что

$$\sup_i \mu(M_i) = \sup_{M \in \mathcal{A}} \mu(M),$$

и положим $M_0 = \bigcup_i M_i$. Тогда $M_0 \in \mathcal{A}$ и одновременно $\mu(M_0) = \sup_{M \in \mathcal{A}} \mu(M)$. Для любого значения параметра $\theta \in Q$ рассмотрим множество

$$M_\theta = \{x \in \mathbb{R}^n \setminus M_0 \mid p_\theta(x) > 0\}.$$

Очевидно, $M_\theta \in \mathcal{A}$ и $M_\theta \cap M_0 = \varnothing$. Поэтому

$$\mu(M_\theta) = \mu(M_\theta \cup M_0) - \mu(M_0) = 0. \quad \square$$

Доказательство теоремы 12.2. Применим лемму 13.1 к выборочному пространству $\mathbb{R}^{nN}$, мере $\mu^N$ и плотности распределения выборки $p_\theta(X)$. Рассмотрим множество $M_0 \subset \mathbb{R}^{nN}$ и счетный набор параметров $\theta_i \in Q$ из этой леммы. Определим с их помощью плотность вероятности

$$p_0(X) = \sum_{i=1}^{\infty} 2^{-i} p_{\theta_i}(X).$$

Она строго положительна на $M_0$. С другой стороны, при каждом $\theta \in Q$ плотность $p_\theta(X)$ обращается в нуль почти всюду вне $M_0$. Не ограничивая общности, можно считать, что она *тождественно* равна нулю вне $M_0$.

*Достаточность.* Пусть $p_\theta(X) = g(T(X), \theta)h(X)$. Рассмотрим функции

$$G(t) = \sum_{i=1}^{\infty} 2^{-i} g(t, \theta_i) \quad \text{и} \quad \rho_\theta(t) = \begin{cases} \dfrac{g(t,\theta)}{G(t)} & \text{при } G(t) \neq 0, \\ 0 & \text{при } G(t) = 0. \end{cases}$$

Очевидно, $p_\theta(X) \equiv \rho_\theta(T(X))p_0(X)$. Отсюда следует, что функция $\rho_\theta(T(X))$ интегрируема с плотностью $p_0(X)$, и интеграл от нее равен единице. Рассмотрим произвольную ограниченную борелевскую функцию $f(X)$. Определим для нее условное математическое ожидание $\bar{f}(T(X)) = \mathsf{E}_0\{f(X)|T(X)\}$. Тогда для любой ограниченной борелевской функции $u\colon \mathbb{R}^k \to \mathbb{R}$ в силу (9.5) будет

$$\int_{\mathbb{R}^{nN}} \bigl[f(X) - \bar{f}(T(X))\bigr] u(T(X)) p_\theta(X)\, \mu^N(dX) =$$

$$= \int_{\mathbb{R}^{nN}} \bigl[f(X) - \bar{f}(T(X))\bigr] u(T(X)) \rho_\theta(T(X)) p_0(X)\, \mu^N(dX) = 0.$$

Поскольку функция $u$ произвольна, отсюда вытекает, что

$$\bar{f}(T(X)) = \mathsf{E}_\theta\{f(X)\,|\,T(X)\}.$$

А в силу произвольности функции $f(X)$ отсюда следует, что условное распределение $P_\theta(dX|T(X))$ совпадает с $P_0(dX|T(X))$ и не зависит от $\theta$.

*Необходимость.* Определим семейства функций

$$q_\theta(X) = \frac{p_\theta(X) + p_0(X)}{2}, \qquad \rho_\theta(X) = \begin{cases} \dfrac{p_\theta(X)}{q_\theta(X)} & \text{при } X \in M_0, \\ 0 & \text{при } X \notin M_0. \end{cases}$$

Очевидно, функция $q_\theta(X)$ строго положительна на $M_0$, а функция $\rho_\theta(X)$ удовлетворяет неравенствам $0 \leqslant \rho_\theta(X) < 2$ и равенству $p_\theta(X) = \rho_\theta(X) q_\theta(X)$. По условию теоремы существует функция $\bar{\rho}_\theta(T(X))$, совпадающая с $\mathsf{E}_\tau\{\rho_\theta(X)|T(X)\}$ при каждом $\tau \in Q$. Тогда она же будет совпадать с $\mathsf{E}\{\rho_\theta(X)|T(X))\}$ по отношению к плотностям $p_0(X)$ и $q_\theta(X)$. Рассмотрим выражение

$$\int_{\mathbb{R}^{nN}} \bigl[\rho_\theta(X) - \bar{\rho}_\theta(T(X))\bigr]^2 q_\theta(X)\, \mu^N(dX) =$$

$$= \int_{\mathbb{R}^{nN}} \bigl[\rho_\theta(X) - \bar{\rho}_\theta(T(X))\bigr] p_\theta(X)\, \mu^N(dX) -$$

$$- \int_{\mathbb{R}^{nN}} \bigl[\rho_\theta(X) - \bar{\rho}_\theta(T(X))\bigr] \bar{\rho}_\theta(T(X)) q_\theta(X)\, \mu^N(dX).$$

В силу свойств условного математического ожидания последние два интеграла обращаются в нуль. Значит, $\rho_\theta(X) = \bar{\rho}_\theta(T(X))$ почти всюду на $M_0$. Подставим это равенство в определение функции $\rho_\theta(X)$. У нас получится

$$\bar{\rho}_\theta(T(X)) = \frac{p_\theta(X)}{q_\theta(X)} = \frac{2p_\theta(X)}{p_\theta(X) + p_0(X)} \quad \Longrightarrow \quad p_\theta(X) = \frac{\bar{\rho}_\theta(T(X))}{2 - \bar{\rho}_\theta(T(X))}\, p_0(X). \quad \square$$

# § 14. Доверительные интервалы

Пусть на пространстве $\mathbb{R}^n$ задано семейство распределений вероятности $\{P_\theta \mid \theta \in (a, b)\}$. До сих пор мы строили *точечные* оценки $\hat{\theta} = T(X)$ для параметра $\theta$. При этом, как правило, вероятность точного совпадения $\hat{\theta}$ с $\theta$ была равна нулю. Спрашивается, а можно ли по выборке $X$ построить такой интервал, который с большой вероятностью содержал бы истинное значение $\theta$?

**Определение.** *Доверительным интервалом* с доверительной вероятностью $1-\varepsilon$ называется такая пара статистик $\underline{\theta} = \underline{\theta}(X)$ и $\bar{\theta} = \bar{\theta}(X)$, что при всех $\theta$ выполняется условие $P_\theta\{\underline{\theta} < \theta < \bar{\theta}\} \geqslant 1 - \varepsilon$.

Таким образом, доверительный интервал $(\underline{\theta}, \bar{\theta})$ — это *случайный* интервал, который содержит истинное значение $\theta$ с вероятностью не меньше $1 - \varepsilon$.

Общая схема построения доверительных интервалов следующая. Предположим, что у нас есть некоторое семейство статистик $T(X, \theta)$, и что для каждой из них известна функция распределения $F_\theta(z) = P_\theta\{T(X, \theta) \leqslant z\}$. Если $F_\theta(z)$ непрерывно зависит от переменной $z$, то $p$-уровень $F_\theta(T(X, \theta))$ равномерно распределен на отрезке $[0, 1]$ (по теореме 4.1). Выберем пару чисел $\varepsilon_1, \varepsilon_2 \geqslant 0$, удовлетворяющих условию $\varepsilon_1 + \varepsilon_2 = \varepsilon$. Тогда с вероятностью $1 - \varepsilon$

$$\varepsilon_2 < F_\theta(T(X, \theta)) < 1 - \varepsilon_1. \tag{14.1}$$

Решим эту систему неравенств относительно переменной $\theta$. Очевидно, любой интервал, содержащий множество всех ее решений, будет доверительным.

Если функция распределения $F_\theta(z)$ разрывна, рассмотрим систему неравенств

$$F_\theta\big(T(X, \theta) - 0\big) < 1 - \varepsilon_1, \qquad F_\theta\big(T(X, \theta)\big) > \varepsilon_2. \tag{14.2}$$

По теореме 4.2 первое из них выполняется с вероятностью не меньше $1-\varepsilon_1$, а второе — с вероятностью не меньше $1-\varepsilon_2$. Значит, вероятность того, что хотя бы одно из них нарушается, не превосходит $\varepsilon_1 + \varepsilon_2 = \varepsilon$, а вероятность того, что они выполняются одновременно, не меньше $1 - \varepsilon$. Поэтому любой интервал, содержащий множество всех решений системы (14.2), будет доверительным.

В большинстве приложений функции $F_\theta(T(X, \theta))$ и $F_\theta(T(X, \theta) - 0)$ оказываются непрерывными и убывающими по $\theta$ (даже если распределение $P_\theta$ дискретно). Тогда границы доверительного интервала $(\underline{\theta}, \bar{\theta})$ можно находить из уравнений:

$$F_{\underline{\theta}}\big(T(X, \underline{\theta}) - 0\big) = 1 - \varepsilon_1, \qquad F_{\bar{\theta}}\big(T(X, \bar{\theta})\big) = \varepsilon_2. \tag{14.3}$$

В случае $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$ полученный таким образом доверительный интервал называется центральным, в случае $\varepsilon_1 = \varepsilon$ доверительный интервал $(\underline{\theta}, +\infty)$ называется левым, а в случае $\varepsilon_2 = \varepsilon$ доверительный интервал $(-\infty, \bar{\theta})$ называется правым.

Размер доверительного интервала зависит от выбора семейства статистик $T(X,\theta)$, а также от пары чисел $\varepsilon_1$, $\varepsilon_2$. Как правило, центральный доверительный интервал имеет длину порядка $1/\sqrt{N}$, где $N$ — объем выборки.

В вышеописанной общей схеме обычно выделяют несколько частных случаев.

**Метод обратной функции** получается, когда статистика $T(X,\theta)$ не зависит от переменной $\theta$ и является состоятельной оценкой для $\theta$. Название метода происходит из традиционного способа его обоснования (которое нам не нужно, поскольку выше мы привели другое, более общее).

**Метод Стьюдента**[1] применяется, когда семейство вероятностных распределений зависит от нескольких параметров, а доверительный интервал нужно построить для одного из них. Обозначим этот выделенный параметр через $\theta$. Предположим, что нам удалось построить такое семейство статистик $T(X,\theta)$, зависящее только от выделенного параметра $\theta$, что его функция распределения $F(z) = P_\theta\{T(X,\theta) \leqslant z\}$ вовсе не зависит от параметров (или зависит лишь от $\theta$). Тогда доверительный интервал, построенный по общей схеме, не будет зависеть от остальных параметров (отличных от $\theta$).

**Асимптотические доверительные интервалы.** Часто удается построить такое семейство статистик $T(X,\theta)$, что его распределение сходится к стандартному нормальному закону $\mathcal{N}(0,1)$. Например, если $\hat\theta(X)$ — асимптотически нормальная оценка $\theta$ с асимптотической дисперсией $\sigma^2(\theta)/N$, то можно взять $T(X,\theta) = \sqrt{N}(\hat\theta(X) - \theta)/\sigma(\theta)$. В этом случае при больших $N$ функцию распределения для $T(X,\theta)$ заменяют на стандартную нормальную функцию распределения $\Phi(z)$, и доверительный интервал для $\theta$ находят из неравенств

$$\frac{\varepsilon}{2} < \Phi(T(X,\theta)) < 1 - \frac{\varepsilon}{2}.$$

Длина построенного таким образом доверительного интервала будет тем меньше, чем меньше асимптотическая дисперсия оценки. Поэтому лучше всего в качестве $\hat\theta(X)$ выбирать асимптотически эффективную оценку (если ее удается вычислить).

[1] Student — псевдоним английского статистика В. Госсета.

**Пример 1.** Пусть $X = (x_1, \ldots, x_N)$ — выборка из нормального закона $\mathcal{N}(\theta, \sigma^2)$ с известной дисперсией $\sigma^2$. Требуется найти доверительный интервал для математического ожидания $\theta$ с доверительной вероятностью $1 - \varepsilon = 0{,}9$.

*Решение.* Состоятельной и несмещенной оценкой для параметра $\theta$ служит выборочное среднее $\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t$. Случайная величина $\bar{x}$ имеет математическое ожидание $\theta$, дисперсию $\sigma^2/N$ и, кроме того, нормально распределена (см. Приложение, п. IV, формулы (5), (6)). Следовательно, статистика $T(X, \theta) = \sqrt{N}(\bar{x} - \theta)/\sigma$ имеет стандартное нормальное распределение. По общей схеме центральный доверительный интервал для $\theta$ находится как решение (относительно переменной $\theta$) системы неравенств

$$\frac{\varepsilon}{2} < \Phi\left(\frac{\sqrt{N}(\bar{x} - \theta)}{\sigma}\right) < 1 - \frac{\varepsilon}{2}.$$

Это решение имеет вид

$$\bar{x} - \Delta\frac{\sigma}{\sqrt{N}} < \theta < \bar{x} + \Delta\frac{\sigma}{\sqrt{N}}, \qquad \text{где} \quad \Delta = \Phi^{-1}\Big(1 - \frac{\varepsilon}{2}\Big) \approx 1{,}65.$$

**Пример 2.** Рассматривается последовательность испытаний Бернулли с двумя исходами $\{0, 1\}$. По выборке $X = (x_1, \ldots, x_N)$ требуется найти доверительный интервал для вероятности успеха $\theta = P(1)$.

*Решение.* Состоятельной и несмещенной оценкой для параметра $\theta$ служит выборочное среднее $\bar{x} = m/N$, где $m = x_1 + \ldots + x_N$ — общее число успехов. Распределение этой оценки биномиальное:

$$P_\theta(m/N) = C_N^m \theta^m (1 - \theta)^{N-m}.$$

В соответствии с (14.3) границы доверительного интервала находятся из уравнений:

$$F_{\underline{\theta}}\Big(\frac{m}{N} - 0\Big) = \sum_{i=0}^{m-1} C_N^i \underline{\theta}^i (1 - \underline{\theta})^{N-i} = 1 - \varepsilon_1,$$

$$F_{\bar{\theta}}\Big(\frac{m}{N}\Big) = \sum_{i=0}^{m} C_N^i \bar{\theta}^{\,i} \big(1 - \bar{\theta}\big)^{N-i} = \varepsilon_2.$$

Если $0 < m < N$, то эти уравнения в явном виде не решаются. Если $m = 0$, то из второго уравнения находим правосторонний доверительный интервал: $\theta \in [0, 1 - \sqrt[N]{\varepsilon})$ с вероятностью $1 - \varepsilon$. А если $m = N$, то из первого уравнения находим левосторонний доверительный интервал: $\theta \in (\sqrt[N]{\varepsilon}, 1]$ с вероятностью $1 - \varepsilon$ (чтобы решить уравнение, используем равенство $F_{\underline{\theta}}(1-0) = 1 - \underline{\theta}^N$).

Оценка $\bar{x} = m/N$ имеет математическое ожидание $\theta$ и дисперсию $\theta(1-\theta)/N$. Распределение статистики $T(X,\theta) = \sqrt{N}(\bar{x} - \theta)\big/\sqrt{\theta(1-\theta)}$ по центральной предельной теореме сходится к стандартному нормальному закону. Поэтому при больших $N$ доверительный интервал можно находить, решая систему неравенств

$$\frac{\varepsilon}{2} < \Phi\left(\frac{\sqrt{N}(\bar{x} - \theta)}{\sqrt{\theta(1-\theta)}}\right) < 1 - \frac{\varepsilon}{2}.$$

**Задача 1.** Докажите, что функция $F_\theta(m/N)$ убывает по $\theta$.
*Подсказка*:

$$\frac{dF_\theta(m/N)}{d\theta} = -(N-m)C_N^m \theta^m (1-\theta)^{N-m-1}.$$

**Пример 3.** Наблюдается выборка $X = (x_1, \ldots, x_N)$ из нормального закона $\mathcal{N}(\theta, \sigma^2)$, в котором $\theta$ и $\sigma^2$ неизвестны. Построить центральный доверительный интервал для $\theta$.

*Решение.* Рассмотрим *статистику Стьюдента*

$$t_{N-1} = \sqrt{N}\,\frac{\bar{x} - \theta}{s}, \tag{14.4}$$

где

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \qquad \text{и} \qquad s^2 = \frac{1}{N-1}\sum_{t=1}^{N}(x_t - \bar{x})^2.$$

Случайные величины $\xi_t = (x_t - \theta)/\sigma$ независимы и имеют стандартное распределение $\mathcal{N}(0,1)$. Сделаем в (14.4) подстановку $x_t = \sigma\xi_t + \theta$:

$$\bar{x} = \sigma\bar{\xi} + \theta, \qquad s^2 = \frac{\sigma^2}{N-1}\sum_{t=1}^{N}(\xi_t - \bar{\xi})^2, \qquad t_{N-1} = \frac{\sqrt{N(N-1)}\cdot\bar{\xi}}{\sqrt{\sum_{t=1}^{N}\big(\xi_t - \bar{\xi}\big)^2}}.$$

Отсюда видно, что распределение статистики $t_{N-1}$ не зависит ни от $\theta$, ни от $\sigma^2$. Позже (во второй части курса) мы докажем, что оно совпадает со стандартным распределением Стьюдента с $N-1$ степенью свободы (см. Приложение, п. VI). Соответствующая функция распределения обозначается $F_{t_{N-1}}$.

Центральный доверительный интервал для $\theta$ находится из системы неравенств

$$\frac{\varepsilon}{2} < F_{t_{N-1}}\left(\sqrt{N}\,\frac{\bar{x}-\theta}{s}\right) < 1-\frac{\varepsilon}{2}.$$

Решая их относительно $\theta$, получаем

$$\bar{x} - \Delta\frac{s}{\sqrt{N}} < \theta < \bar{x} + \Delta\frac{s}{\sqrt{N}}, \qquad \text{где} \quad \Delta = F^{-1}_{t_{N-1}}\left(1-\frac{\varepsilon}{2}\right).$$

Квантиль $\Delta$ находится из таблиц для функции распределения $F_{t_{N-1}}$ или с помощью вероятностного калькулятора.

**Пример 4.** Наблюдается выборка $X = (x_1, \ldots, x_N)$ из нормального закона $\mathcal{N}(\theta, \sigma^2)$, в котором $\theta$ и $\sigma^2$ неизвестны. Построить центральный доверительный интервал для $\sigma^2$.

*Решение.* Рассмотрим статистику

$$\chi^2_{N-1} = \frac{N-1}{\sigma^2}s^2 = \frac{1}{\sigma^2}\sum_{t=1}^{N}(x_t-\bar{x})^2 = \sum_{t=1}^{N}\bigl(\xi_t-\bar{\xi}\bigr)^2,$$

где $\xi_t$ — случайные величины из предыдущего примера. Ее распределение не зависит от параметров $\theta$ и $\sigma^2$. Во второй части курса мы докажем, что оно совпадает с распределением «хи-квадрат» с $N-1$ степенью свободы (см. Приложение, п. VI). Поэтому центральный доверительный интервал для $\sigma^2$ можно находить из системы

$$\frac{\varepsilon}{2} < F_{\chi^2_{N-1}}\left(\frac{N-1}{\sigma^2}s^2\right) < 1-\frac{\varepsilon}{2}.$$

Решая ее, получаем

$$\frac{(N-1)s^2}{F^{-1}_{\chi^2_{N-1}}(1-\varepsilon/2)} < \sigma^2 < \frac{(N-1)s^2}{F^{-1}_{\chi^2_{N-1}}(\varepsilon/2)}. \tag{14.5}$$

**Задача 2.** Докажите, что доверительный интервал (14.5) имеет длину порядка $1/\sqrt{N}$.

# Глава 2. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

## § 15. Основные понятия

Пусть на пространстве $\mathbb{R}^n$ имеется семейство вероятностных мер $\{\, P_\theta \mid \theta \in Q \,\}$, и множество параметров $Q$ разбито на несколько частей: $Q = Q_0 \sqcup Q_1 \sqcup \ldots \sqcup Q_{K-1}$. Это разбиение определяет набор *гипотез*: гипотеза $H_i$ состоит в том, что истинное значение параметра $\theta$ принадлежит множеству $Q_i$. Если множество $Q_i$ содержит только один элемент, то гипотеза $H_i$ называется простой; в противном случае она называется сложной. Задача статистической проверки гипотез состоит в построении статистики $d(X)$, сопоставляющей каждой выборке $X$ наиболее вероятную гипотезу. Такая статистика называется *решающим правилом (критерием, тестом).*

Рассматривают две разновидности решающих правил. Если значение функции $d(X)$ однозначно определяется по выборке $X$, то это обычное (нерандомизированное) решающее правило. А если при фиксированном $X$ значение $d(X)$ определяется не однозначно, а само является случайной величиной, то это рандомизированное решающее правило. Для него вероятности $P\{d(X) = H_i\} = \varphi_i(X)$ называются критическими функциями. Очевидно, обычное решающее правило — это частный случай рандомизированного.

Чаще всего изучается ситуация, когда множество $Q$ разбито всего на две части: $Q = Q_0 \sqcup Q_1$. Тогда есть две гипотезы: гипотеза $H_0$ состоит в том, что $\theta \in Q_0$ (нулевая гипотеза), а гипотеза $H_1$ состоит в том, что $\theta \in Q_1$ (альтернатива).

В данной ситуации любое рандомизированное решающее правило — это случайная функция $d(X)$, принимающая значение $H_1$ с вероятностью $\varphi(X)$ и значение $H_0$ с вероятностью $1 - \varphi(X)$, где $\varphi(X)$ — произвольная (борелевская) критическая функция, принимающая значения в отрезке $[0, 1]$.

Ошибкой I рода называется выбор гипотезы $H_1$ при условии, что на самом деле верна $H_0$. Вероятность ошибки первого рода есть

$$\alpha(\theta) = P_\theta\{d(X) = H_1\} = \int_{\mathbb{R}^{nN}} \varphi(X)\, P_\theta(dX), \qquad \theta \in Q_0.$$

Ошибкой II рода называется принятие гипотезы $H_0$, когда на самом деле верна $H_1$. Вероятность ошибки второго рода есть

$$\beta(\theta) = P_\theta\{d(X) = H_0\} = \int_{\mathbb{R}^{nN}} \bigl(1 - \varphi(X)\bigr)\, P_\theta(dX), \qquad \theta \in Q_1.$$

Мощность решающего правила $d(X)$ — это вероятность правильного принятия альтернативы $H_1$. Она обозначается $\omega(\theta)$ и равна

$$\omega(\theta) = P_\theta\{d(X) = H_1\} = \int_{\mathbb{R}^{nN}} \varphi(X)\, P_\theta(dX) = 1 - \beta(\theta), \qquad \theta \in Q_1.$$

Разумно считать решающее правило тем лучше, чем меньше для него вероятности ошибок первого и второго рода. Однако в большинстве случаев уменьшение вероятности ошибки одного рода ведет к увеличению вероятности ошибки другого рода, и возникает задача достижения некоторого баланса между ними. На практике обычно гипотезы $H_0$ и $H_1$ бывают неравноправны, и ошибки первого рода менее желательны, чем ошибки второго рода. В связи с этим Нейман и Пирсон предложили следующий принцип оптимальности решающего правила.

**Принцип оптимальности.** Рандомизированное решающее правило следует выбирать так, чтобы вероятность ошибки первого рода не превосходила некоторого наперед заданного малого числа $\varepsilon > 0$, а вероятность ошибки второго рода была минимально возможной:

$$\sup_{\theta\in Q_0} \alpha(\theta) \leqslant \varepsilon, \qquad \sup_{\theta\in Q_1} \beta(\theta) \to \min \quad \Bigl[\Leftrightarrow \inf_{\theta\in Q_1} \omega(\theta) \to \max\Bigr].$$

Число $\varepsilon$ называется *уровнем значимости* теста, а само оптимальное решающее правило — решающим правилом Неймана — Пирсона.

Если для любого $\theta \in Q_1$ вероятность ошибки второго рода $\beta(\theta)$ стремится к нулю при $N \to \infty$, то тест называется *состоятельным*. А если вероятность принятия альтернативы при любом $\theta \in Q_1$ выше, чем при любом $\theta \in Q_0$ (другими словами, выполняется неравенство $\sup_{\theta\in Q_0} \alpha(\theta) \leqslant \inf_{\theta\in Q_1} \omega(\theta)$), то тест называется *несмещенным*.

Название «уровень значимости» теста кажется несколько противоестественным: получается, что при прочих равных условиях тест тем лучше, чем меньше его уровень значимости. Но тут ничего не поделаешь; этот термин общепринят, и не нам его менять (впрочем, Боровков в [1] все-таки отважился назвать уровнем значимости не $\varepsilon$, а $1 - \varepsilon$).

# § 16. Решающее правило Неймана — Пирсона

Сами Нейман и Пирсон в 1933 году построили оптимальное решающее правило в случае двух простых гипотез. Предположим, что на $\mathbb{R}^n$ задана $\sigma$-конечная борелевская мера $\mu$ и две плотности вероятности $p_0(x)$, $p_1(x)$ по отношению к $\mu$. Будем считать, что множество параметров $Q$ содержит всего два элемента, $\theta_0$ и $\theta_1$, которым отвечают плотности $p_0(x)$ и $p_1(x)$. Гипотеза $H_0$ состоит в том, что мы наблюдаем случайную величину с плотностью распределения $p_0(x)$, а гипотеза $H_1$ состоит в том, что эта случайная величина имеет плотность $p_1(x)$.

Пусть $X = (x_1, \ldots, x_N)$ — выборка из распределения с плотностью $p_i(x)$, где $i = 0, 1$. Тогда *отношением правдоподобия* называется статистика

$$L(X) = \frac{p_1(X)}{p_0(X)}.$$

**Теорема 16.1.** *В случае двух простых гипотез для любого числа $\varepsilon \in (0,1)$ существуют такие числа $\Delta \geqslant 0$ и $\varkappa \in [0,1]$, что у решающего правила с критической функцией*

$$\varphi(X) = P\{d(X) = H_1\} = \begin{cases} 0, & \text{если } L(X) < \Delta, \\ \varkappa, & \text{если } L(X) = \Delta, \\ 1, & \text{если } L(X) > \Delta \end{cases} \tag{16.1}$$

*вероятность ошибки первого рода равна $\varepsilon$. Это решающее правило имеет максимальную мощность среди всех решающих правил с уровнем значимости $\varepsilon$.*

Доказательство. Пусть $P_0$ — распределение вероятностей с плотностью $p_0(X)$, и $F_0(z) = P_0\{L(X) \leqslant z\}$ — функция распределения отношения правдоподобия. Тогда вероятность ошибки первого рода для теста с критической функцией $\varphi(X)$ равна

$$\begin{aligned} \alpha &= P_0\{L(X) > \Delta\} + \varkappa P_0\{L(X) = \Delta\} = \\ &= \big(1 - F_0(\Delta)\big) + \varkappa\big(F_0(\Delta) - F_0(\Delta - 0)\big). \end{aligned}$$

При $\varkappa = 0$ она принимает значение $1 - F_0(\Delta)$, а при $\varkappa = 1$ она равна $1 - F_0(\Delta - 0)$. Пусть $\Delta = \sup\{z \mid F_0(z) < 1 - \varepsilon\}$ — квантиль уровня $1 - \varepsilon$

для функции распределения $F_0(z)$. Тогда $F_0(\Delta - 0) \leqslant 1 - \varepsilon \leqslant F_0(\Delta)$ или, что то же самое, $1 - F_0(\Delta) \leqslant \varepsilon \leqslant 1 - F_0(\Delta - 0)$. Очевидно, для этого $\Delta$ найдется такое $\varkappa \in [0, 1]$, при котором $\alpha = \varepsilon$.

Мощность построенного решающего правила равна

$$\omega = \int_{\mathbb{R}^{nN}} \varphi(X) p_1(X)\, \mu^N(dX).$$

Рассмотрим любую другую критическую функцию $\varphi'(X)$, для которой

$$\alpha' = \int_{\mathbb{R}^{nN}} \varphi'(X) p_0(X)\, \mu^N(dX) \leqslant \varepsilon,$$

$$\omega' = \int_{\mathbb{R}^{nN}} \varphi'(X) p_1(X)\, \mu^N(dX).$$

Из определения $\varphi(X)$ видно, что функция

$$G(X) = \big(\varphi(X) - \varphi'(X)\big)\big(p_1(X) - \Delta p_0(X)\big)$$

неотрицательна. Значит,

$$\int_{\mathbb{R}^{nN}} G(X)\, \mu^N(dX) = \omega - \omega' - \Delta(\varepsilon - \alpha') \geqslant 0,$$

откуда следует, что $\omega \geqslant \omega'$. $\square$

**Следствие 16.2.** *Если функция распределения*

$$F_0(z) = P_0\{L(X) \leqslant z\}$$

*непрерывна, то существует нерандомизированное решающее правило Неймана — Пирсона вида*

$$d(X) = \begin{cases} H_0, & \text{если} \quad L(X) \leqslant \Delta, \\ H_1, & \text{если} \quad L(X) > \Delta, \end{cases} \qquad \text{где} \quad \Delta \in F_0^{-1}(1 - \varepsilon). \tag{16.2}$$

Доказательство. В этом случае число $\varkappa$ можно взять любым, например $\varkappa = 0$. $\square$

В примерах довольно часто удается подобрать такую возрастающую функцию $g$, что выражение $g(L(X))$ выглядит проще, чем $L(X)$.

Тогда решающее правило (16.2) записывают в равносильной форме

$$d(X) = \begin{cases} H_0, & \text{если } \ g(L(X)) \leqslant \delta, \\ H_1, & \text{если } \ g(L(X)) > \delta, \end{cases} \qquad \text{где } \ \delta = g(\Delta).$$

**Пример.** Пусть имеются два нормальных распределения с одинаковыми дисперсиями $\sigma^2$, но разными математическими ожиданиями $\theta_0 < \theta_1$. Для выборки $X = (x_1, \ldots, x_N)$ нужно построить решающее правило Неймана — Пирсона с уровнем значимости $\varepsilon$.

*Решение.* Отношение правдоподобия в данном случае имеет вид

$$L(X) = \frac{p_1(X)}{p_0(X)} = \prod_{t=1}^{N} \frac{e^{-(x_t-\theta_1)^2/2\sigma^2}}{e^{-(x_t-\theta_0)^2/2\sigma^2}} = \\ = \exp\left\{ \frac{N(\theta_1 - \theta_0)}{\sigma^2}\, \bar{x} \,-\, \frac{N(\theta_1^2 - \theta_0^2)}{2\sigma^2} \right\}.$$

Очевидно, $L(X)$ является возрастающей функцией от $\bar{x}$. Поэтому для каждого $\Delta > 0$ существует единственное число $\delta$, при котором условие $L(X) > \Delta$ равносильно $\bar{x} > \delta$. Заметим, что при истинном математическом ожидании $\theta_0$ выборочное среднее $\bar{x}$ имеет нормальное распределение $\mathcal{N}(\theta_0, \sigma^2/N)$. Поэтому величина $y = \sqrt{N}(\bar{x} - \theta_0)/\sigma$ имеет стандартное нормальное распределение $\mathcal{N}(0,1)$. А неравенство $\bar{x} > \delta$ равносильно $y > \zeta$, где $\zeta = \sqrt{N}(\delta - \theta_0)/\sigma$.

В решающем правиле Неймана — Пирсона вероятность ошибки первого рода должна равняться $\varepsilon$:

$$\alpha = P_0\{L(X) > \Delta\} = P_0\{\bar{x} > \delta\} = P_0\{y > \zeta\} = \varepsilon. \tag{16.3}$$

С другой стороны, $P_0\{y > \zeta\} = 1 - \Phi(\zeta)$. Следовательно, $1 - \Phi(\zeta) = \varepsilon$. Отсюда мы находим квантиль $\zeta = \Phi^{-1}(1-\varepsilon)$, критическое значение $\delta = \theta_0 + \zeta\sigma/\sqrt{N}$, и в итоге получаем решающее правило

$$\begin{cases} \theta = \theta_0, & \text{если } \ \bar{x} \leqslant \delta = \theta_0 + \zeta\dfrac{\sigma}{\sqrt{N}}, \\ \theta = \theta_1, & \text{если } \ \bar{x} > \delta = \theta_0 + \zeta\dfrac{\sigma}{\sqrt{N}}, \end{cases} \qquad \text{где} \quad \zeta = \Phi^{-1}(1-\varepsilon). \tag{16.4}$$

Докажем его состоятельность. Если $\theta = \theta_1$, то случайная величина $z = \sqrt{N}(\bar{x} - \theta_1)/\sigma$ имеет нормальное распределение $\mathcal{N}(0,1)$. Поэтому вероятность ошибки второго рода

$$P_1\{\bar{x} \leqslant \delta\} = P_1\big\{z \leqslant \sqrt{N}(\theta_0 - \theta_1)/\sigma + \zeta\big\} = \Phi\big(\sqrt{N}(\theta_0 - \theta_1)/\sigma + \zeta\big\}$$

стремится к нулю при $N \to \infty$.

**Вопрос.** Как можно объяснить тот факт, что критерий (16.4) зависит лишь от $\bar{x}$?

# § 17. Проверка простой гипотезы против сложной альтернативы

Пусть на $\mathbb{R}^n$ задано семейство вероятностных мер $\{\, P_\theta \mid \theta \in Q \,\}$, и в множестве параметров $Q$ фиксировано одно значение $\theta_0$. Требуется проверить простую гипотезу $H_0\colon\ \theta = \theta_0$ против альтернативы общего вида $H_1\colon\ \theta \neq \theta_0$.

Обычно это делают так. Выбирают какую-нибудь скалярную статистику $T(X)$, для которой при $\theta = \theta_0$ известна функция распределения $F(z) = P_{\theta_0}\{T(X) \leqslant z\}$. В области изменения $T(X)$ находят такой отрезок $[\Delta_1, \Delta_2]$, что $P_{\theta_0}\{T(X) \in [\Delta_1, \Delta_2]\} \geqslant 1 - \varepsilon$. Для него выписывают решающее правило

$$\begin{cases} \theta = \theta_0, & \text{если } \ T(X) \in [\Delta_1, \Delta_2], \\ \theta \neq \theta_0, & \text{если } \ T(X) \notin [\Delta_1, \Delta_2]. \end{cases} \tag{17.1}$$

По построению оно имеет уровень значимости $\varepsilon$. После этого для каждого $\theta \neq \theta_0$ проверяют, будет ли вероятность $P_\theta\{T(X) \in [\Delta_1, \Delta_2]\}$ стремиться к нулю при возрастании объема выборки $X$. Если это так, то решающее правило (17.1) состоятельно.

Предположим, что функция распределения $F(z) = P_{\theta_0}\{T(X) \leqslant z\}$ непрерывна. Тогда решающее правило (17.1) можно записать в равносильной форме

$$\begin{cases} \theta = \theta_0, & \text{если } \ F(T(X)) \in [\delta_1, \delta_2], \\ \theta \neq \theta_0, & \text{если } \ F(T(X)) \notin [\delta_1, \delta_2], \end{cases} \qquad \text{где} \quad \delta_i = F(\Delta_i). \tag{17.2}$$

Напомним, что случайная величина $F(T(X))$ называется *p-уровнем* статистики $T(X)$. В силу теоремы 4.1 она равномерно распределена на отрезке $[0,1]$ (при $\theta = \theta_0$). Поэтому

$$\delta_2 - \delta_1 = P_{\theta_0}\{F(T(X)) \in [\delta_1, \delta_2]\} \geqslant 1 - \varepsilon.$$

Если функция распределения $F(z)$ разрывна, то вместо решающего правила (17.2) можно использовать следующее:

$$\begin{cases} \theta = \theta_0, & \text{если } F(T(X)) > \delta_1 \text{ и } F(T(X)-0) < \delta_2, \\ \theta \neq \theta_0, & \text{если } F(T(X)) \leqslant \delta_1 \text{ или } F(T(X)-0) \geqslant \delta_2, \end{cases} \tag{17.3}$$

где $\delta_2 - \delta_1 = 1 - \varepsilon$. По теореме 4.2 неравенство $F(T(X)) > \delta_1$ выполняется с вероятностью не меньше $1 - \delta_1$, а неравенство $F(T(X)-0) < \delta_2$ выполняется с вероятностью не меньше $\delta_2$. Вероятность же того, что хотя бы одно из них нарушается, не превосходит $\delta_1 + (1 - \delta_2) = \varepsilon$. Это доказывает, что последний тест имеет уровень значимости $\varepsilon$.

В приложениях часто рассматривают ситуацию, когда статистика $T(X)$ неотрицательна и характеризует величину отклонения некоторой статистической оценки $\hat{\theta}$ от $\theta_0$. Например, $T(X)$ может быть расстоянием $|\hat{\theta} - \theta_0|$ или квадратом расстояния $|\hat{\theta} - \theta_0|^2$. В этом случае обычно полагают $\delta_1 = 0,\ \delta_2 = 1 - \varepsilon$, и записывают критерий (17.2) в виде

$$\begin{cases} \theta = \theta_0, & \text{если } 1 - F(T(X)) \geqslant \varepsilon, \\ \theta \neq \theta_0, & \text{если } 1 - F(T(X)) < \varepsilon, \end{cases} \tag{17.4}$$

При этом $p$-уровнем называют не $F(T(X))$, а разность $1 - F(T(X))$ (которая, очевидно, тоже равномерно распределена на отрезке $[0,1]$). Для статистики $T(X) \geqslant 0$ выполняются равенства $F(0) = 0$ и $F(+\infty) = 1$. Поэтому малым значениям $T(X)$ соответствуют $p$-уровни, близкие к единице, а большим значениям $T(X)$ отвечают $p$-уровни, близкие к нулю. При использовании критерия (17.4) принято говорить, что если $p$-уровень оказался выше уровня значимости ($1 - F(T(X)) \geqslant \varepsilon$), то статистика $T(X)$ незначимо отклонилась от нуля, и гипотезу $\theta = \theta_0$ можно принять. Если же $p$-уровень оказался ниже уровня значимости ($1 - F(T(X)) < \varepsilon$), то статистика $T(X)$ значимо отклонилась от нуля, и поэтому гипотезу $\theta = \theta_0$ следует отвергнуть.

Приведем еще один способ истолкования $p$-уровня. Пусть $X$ и $Y$ — две независимые выборки одинаковой мощности из распределения $P_{\theta_0}$, причем выборка $X$ фиксирована. Тогда

$$P_{\theta_0}\{T(Y) > T(X)\} = 1 - F(T(X)).$$

Другими словами, $p$-уровень выборки $X$ — это вероятность события $T(Y) > T(X)$ при $\theta = \theta_0$. И если эта вероятность оказалась меньше $\varepsilon$, то мы считаем, что статистика $T(X)$ приняла слишком большое значение (значимо отклонилась от нуля), чтобы можно было поверить в гипотезу $\theta = \theta_0$.

В заключение еще раз подчеркнем, что состоятельность критериев (17.1) – (17.4) *a priori* ничем не гарантирована, и ее всегда нужно проверять отдельно.

**Пример 1.** Пусть $X = (x_1, \ldots, x_N)$ — выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$ с известной дисперсией $\sigma^2$. Построить критерий для проверки гипотезы $a = a_0$.

*Решение.* Известно, что при $a = a_0$ статистика

$$T(X) = \frac{1}{\sigma\sqrt{N}} \sum_{t=1}^{N} (x_t - a_0) = \frac{\sqrt{N}}{\sigma}\,(\bar{x} - a_0)$$

имеет стандартное нормальное распределение $\mathcal{N}(0, 1)$ (это следует из формул (5), (6) в Приложении, п. IV). Подберем такое число $\Delta$, при котором $P\{|T(X)| \leqslant \Delta\} = 1 - \varepsilon$. Очевидно, это $\Delta = \Phi^{-1}(1 - \varepsilon/2)$. Соответствующий критерий с уровнем значимости $\varepsilon$ имеет вид

$$\begin{cases} a = a_0, & \text{если } |T(X)| \leqslant \Delta, \\ a \neq a_0, & \text{если } |T(X)| > \Delta, \end{cases} \qquad \text{где} \quad \Delta = \Phi^{-1}\Big(1 - \frac{\varepsilon}{2}\Big).$$

Проверим его состоятельность. Если $a \neq a_0$, то стандартное нормальное распределение будет иметь статистика

$$\sqrt{N}(\bar{x} - a)/\sigma = T(X) - \sqrt{N}(a - a_0)/\sigma.$$

Поэтому при $N \to \infty$

$$P\big\{|T(X)| \leqslant \Delta\big\} \leqslant P\left\{\frac{\sqrt{N}\,|\bar{x} - a|}{\sigma} \geqslant \frac{\sqrt{N}\,|a - a_0|}{\sigma} - \Delta\right\} \to 0.$$

**Пример 2.** Пусть $X = (x_1, \ldots, x_N)$ — выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$ с неизвестными $a$ и $\sigma$. Построить критерий для проверки гипотезы $\sigma = \sigma_0$.

*Решение.* Во второй части курса будет доказано, что при $\sigma = \sigma_0$ статистика

$$\chi^2(X) = \frac{1}{\sigma_0^2} \sum_{t=1}^{N} (x_t - \bar{x})^2$$

имеет распределение $\chi^2_{N-1}$ (см. Приложение, п. VI). Определим квантили $\Delta_1$ и $\Delta_2$ равенствами

$$F_{\chi^2_{N-1}}(\Delta_1) = \frac{\varepsilon}{2}, \qquad F_{\chi^2_{N-1}}(\Delta_2) = 1 - \frac{\varepsilon}{2}.$$

Тогда при $\sigma = \sigma_0$ вероятность события $\chi^2(X) \in [\Delta_1, \Delta_2]$ будет равна $1 - \varepsilon$, а вероятность события $\chi^2(X) \notin [\Delta_1, \Delta_2]$ будет равна $\varepsilon$. Соответствующий критерий имеет вид

$$\begin{cases} \sigma = \sigma_0, & \text{если } \chi^2(X) \in [\Delta_1, \Delta_2], \\ \sigma \neq \sigma_0, & \text{если } \chi^2(X) \notin [\Delta_1, \Delta_2]. \end{cases}$$

Проверим состоятельность этого критерия. По определению, $\chi^2_n = \xi_1^2 + \ldots + \xi_n^2$, где случайные величины $\xi_i$ независимы и имеют стандартное нормальное распределение $\mathcal{N}(0, 1)$. Из центральной предельной теоремы вытекает, что при $N \to \infty$ сумма $\sum_{t=1}^{N} (x_t - \bar{x})^2$ с подавляющей вероятностью попадает в окрестность точки $(N-1)\sigma^2$ размера порядка $\sqrt{N}$, а условие $\chi^2(X) \in [\Delta_1, \Delta_2]$ задает для той же суммы интервал, лежащий в окрестности точки $(N-1)\sigma_0^2$ размера порядка $\sqrt{N}$. Поэтому при $\sigma \neq \sigma_0$ вероятность события $\chi^2(X) \in [\Delta_1, \Delta_2]$ стремится к нулю, что доказывает состоятельность построенного критерия.

# § 18. Критерии согласия

Предположим, что на вещественной оси есть какая-то функция распределения $F(x)$, и мы наблюдаем выборку $X = (x_1, \ldots, x_N)$ из этого распределения. Требуется решить, совпадает ли $F(x)$ с некоторой другой (заранее заданной) функцией распределения $F_0(x)$. Иначе говоря, нам нужно проверить простую гипотезу $H_0 : F(x) \equiv F_0(x)$ против альтернативы общего вида $H_1 : F(x) \not\equiv F_0(x)$. При такой постановке задачи $H_0$ называется гипотезой согласия (потому что она утверждает,

что выборка $X$ «согласуется» с распределением $F_0(x)$), а соответствующее ей решающее правило называется критерием согласия. Наиболее популярны два критерия согласия: $\chi^2$-критерий Пирсона и критерий Колмогорова.

**$\chi^2$-критерий Пирсона.** Произвольным образом разобьем вещественную прямую на $m$ частей точками

$$-\infty = b_0 < b_1 < \ldots < b_{m-1} < b_m = +\infty.$$

Положим $\Delta_k = (b_{k-1}, b_k]$. Вероятность того, что случайная величина $x$ с функцией распределения $F_0(x)$ попадет в $\Delta_k$, равна

$$p_k = F_0(b_k) - F_0(b_{k-1}).$$

Ниже мы будем предполагать, что все $p_k > 0$. Для каждой выборки $X = (x_1, \ldots, x_N)$ обозначим через $\nu_k$ число выборочных значений $x_i$, попавших в $\Delta_k$, а через $\hat{p}_k$ — эмпирические вероятности $\hat{p}_k = \nu_k/N$.

**Определение.** *$\chi^2$-статистика Пирсона* — это

$$\chi^2(X) = \sum_{k=1}^{m} \frac{(\nu_k - Np_k)^2}{Np_k} = N \sum_{k=1}^{m} \frac{(\hat{p}_k - p_k)^2}{p_k}. \tag{18.1}$$

**Теорема 18.1.** *Если $X = (x_1, \ldots, x_N)$ — выборка из распределения $F_0(x)$, то при $N \to \infty$ распределение $\chi^2$-статистики Пирсона сходится к распределению случайной величины $\chi^2_{m-1}$* (распределению «хи-квадрат» с $m-1$ степенями свободы).

Доказательство этой теоремы будет изложено позднее, во второй части курса.

По определению, $\chi^2$-критерий согласия Пирсона имеет вид

$$\begin{cases} F(x) \equiv F_0(x), & \text{если } \chi^2(X) \leqslant \Delta, \\ F(x) \not\equiv F_0(x), & \text{если } \chi^2(X) > \Delta, \end{cases} \qquad \Delta = F^{-1}_{\chi^2_{m-1}}(1-\varepsilon),$$

где $F_{\chi^2_{m-1}}(x)$ — функция распределения случайной величины $\chi^2_{m-1}$.

Из теоремы 18.1 вытекает, что вероятность ошибки первого рода $P\{\chi^2(X) > \Delta \mid H_0\}$ сходится к $\varepsilon$ при $N \to \infty$. Мы всегда можем снизить эту вероятность до приемлемого уровня за счет выбора достаточно малого $\varepsilon$.

С другой стороны, допустим, что выборка $X$ подчиняется распределению $F(x)$, отличному от $F_0(x)$. Тогда числа $q_k = F(b_k) - F(b_{k-1})$ скорее всего будут отличаться от $p_k = F_0(b_k) - F_0(b_{k-1})$. По закону больших чисел эмпирические вероятности $\hat{p}_k$ почти наверное будут сходиться к $q_k$ при $N \to \infty$. Поэтому отношение $\chi^2(X)/N$ будет сходиться к числу $c = \sum_k (q_k - p_k)^2/p_k > 0$, сама случайная величина $\chi^2(X)$ будет неограниченно возрастать, и выбор будет делаться в пользу гипотезы $H_1\colon\ F(x) \not\equiv F_0(x)$ (иначе говоря, рассматриваемый критерий является состоятельным).

К сожалению, вероятность ошибки второго рода стремится к нулю не равномерно по отношению к распределению $F(x) \not\equiv F_0(x)$. Чем меньше отличается $F(x)$ от $F_0(x)$, тем больше нужно брать размер выборки, чтобы гарантировать малость этой вероятности. А в случае, когда распределение $F(x)$ может быть сколь угодно близко к $F_0(x)$, мы ни при каком $N$ не можем утверждать, что вероятность ошибки второго рода мала.

**Критерий Колмогорова.** Пусть $X = (x_1, \ldots, x_N)$ — выборка из распределения с непрерывной функции распределения $F_0(x)$. В силу теоремы 4.1 случайные величины $y_i = F_0(x_i)$ равномерно распределены на отрезке $[0, 1]$. Рассмотрим выборку $Y = (y_1, \ldots, y_N)$. Построим для нее выборочную функцию распределения

$$\Psi_N(y) = \frac{\#\{\, y_i \mid y_i \leqslant y \,\}}{N} = \frac{1}{N} \sum_{i=1}^{N} \eta(y - y_i),$$

где $\eta(x)$ — функция Хевисайда (обращающаяся в нуль при $x < 0$ и равная единице при $x \geqslant 0$).

**Определение.** *Статистика (расстояние) Колмогорова — это*

$$D_N(X) = \sup_{0 \leqslant y \leqslant 1} \bigl|\Psi_N(y) - y\bigr|. \tag{18.2}$$

Очевидно, распределение статистики $D_N(X)$ не зависит от распределения $F_0(x)$ (при условии, что оно непрерывно). Из сильной состоятельности выборочной функции распределения (теорема 3.1) вытекает, что $D_N(X) \to 0$ с вероятностью 1.

**Задача 1.** Докажите справедливость равенств

$$D_N(X) = \sup_x \bigl|F_N(x) - F_0(x)\bigr|,$$

$$D_N(X) = \max_{1\leqslant i\leqslant N} \left\{ \Bigl|y_{(i)} - \frac{i}{N}\Bigr|, \Bigl|y_{(i)} - \frac{i-1}{N}\Bigr| \right\},$$

в которых $F_N(x)$ — выборочная функция распределения для выборки $X = (x_1, \ldots, x_N)$, а $(y_{(1)}, \ldots, y_{(N)})$ — вариационный ряд для выборки $Y = (y_1, \ldots, y_N)$, где $y_i = F_0(x_i)$.

**Теорема 18.2.** *Если выборка $X = (x_1, \ldots, x_N)$ отвечает непрерывной функциии распределения $F_0(x)$, то распределение статистики $\sqrt{N}D_N(X)$ сходится к распределению Колмогорова*:

$$P\bigl\{\sqrt{N}D_N(X) \leqslant z\bigr\} \longrightarrow K(z) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2z^2}, \qquad z > 0.$$

Эта теорема приводится без доказательства.

По определению, критерий согласия Колмогорова имеет вид

$$\begin{cases} F(x) \equiv F_0(x), & \text{если } \sqrt{N}D_N(X) \leqslant K^{-1}(1-\varepsilon), \\ F(x) \not\equiv F_0(x), & \text{если } \sqrt{N}D_N(X) > K^{-1}(1-\varepsilon). \end{cases}$$

Из теоремы 18.2 следует, что вероятность ошибки первого рода для этого критерия сходится к $\varepsilon$ при $N \to \infty$.

**Задача 2.** Объясните, почему критерий Колмогорова состоятелен.

# § 19. Критерий отношения правдоподобия

Пусть на пространстве $\mathbb{R}^n$ задана $\sigma$-конечная мера $\mu$ и семейство плотностей вероятностей $\{p_\theta(x) \mid \theta \in Q\}$ по отношению к $\mu$. Предположим, что множество $Q$ разбито на две части $Q_0 \subset Q$ и $Q_1 = Q \setminus Q_0$. *Отношением правдоподобия* для проверки двух гипотез $H_0\colon\ \theta \in Q_0$ и $H_1\colon\ \theta \in Q_1$ называется статистика

$$\Lambda(X) = \frac{\sup\{p_\theta(X) \mid \theta \in Q\}}{\sup\{p_\theta(X) \mid \theta \in Q_0\}}, \tag{19.1}$$

где $p_\theta(X) = p_\theta(x_1)\dots p_\theta(x_N)$ — это плотность распределения выборки $X = (x_1, \dots, x_N)$. Очевидно, $\Lambda(X) \geqslant 1$.

По определению, критерий отношения правдоподобия имеет вид

$$\begin{cases} \theta \in Q_0, & \text{если } \ \Lambda(X) \leqslant \Delta, \\ \theta \notin Q_0, & \text{если } \ \Lambda(X) > \Delta, \end{cases} \tag{19.2}$$

где критическое значение $\Delta > 1$ выбирается так, чтобы критерий имел заранее заданный уровень значимости $\varepsilon$:

$$\alpha(\theta) = P_\theta\{\Lambda(X) > \Delta\} = \int_{\Lambda(X)>\Delta} p_\theta(X)\,\mu^N(dX) \ \leqslant \ \varepsilon$$

для всех $\theta \in Q_0$.

Оказывается, если семейство плотностей $p_\theta(x)$ удовлетворяет некоторым условиям регулярности (сформулированным в следующей теореме 19.1), а множество $Q_0$ является гладким подмногообразием в области $Q$, то при $\theta \in Q_0$ распределение статистики $2\ln\Lambda(X)$ сходится к распределению $\chi^2_s$, причем число степеней свободы $s$ равно коразмерности множества $Q_0$ в $Q$. Отсюда вытекает, что при больших $N$

$$P_\theta\{\Lambda(X) > \Delta\} \approx P_\theta\{\chi^2_s > 2\ln\Delta\} = 1 - F_{\chi^2_s}(2\ln\Delta).$$

Приравнивая последнее выражение к $\varepsilon$, находим критическое значение $\Delta = e^{\delta/2}$, где число $\delta = 2\ln\Delta$ определяется условием $F_{\chi^2_s}(\delta) = 1 - \varepsilon$.

Сформулируем соответствующую теорему об асимптотике распределения статистики $2\ln\Lambda(X)$.

**Теорема 19.1.** *Предположим, что $Q$ — открытая область в $\mathbb{R}^m$, семейство плотностей вероятностей $\{\,p_\theta(x) \mid \theta \in Q\,\}$ (по отношению к некоторой $\sigma$-конечной борелевской мере $\mu$ на $\mathbb{R}^n$) удовлетворяет трем условиям регулярности*

*$R1$) выражение $\int_{\mathbb{R}^n} p_\theta(x)\,\mu(dx)$ можно дважды дифференцировать по $\theta$ под знаком интеграла;*

*$R2$) информационная матрица Фишера*

$$\mathcal{I}(\theta) = \big(i_{kl}(\theta)\big)_{k,l=1}^m, \qquad i_{kl}(\theta) = \mathsf{E}_\theta\left\{\frac{\partial \ln p_\theta(x)}{\partial\theta_k}\,\frac{\partial \ln p_\theta(x)}{\partial\theta_l}\right\}$$

*положительно определена при всех $\theta \in Q$;*

*R*3) *существуют такие функция* $H(x)$ *и константа* $M$, *что при всех* $\theta \in Q$

$$\left|\frac{\partial^3 \ln p_\theta(x)}{\partial\theta_j \partial\theta_k \partial\theta_l}\right| \leqslant H(x), \quad \mathsf{E}_\theta\{H(x)\} \leqslant M, \qquad j,k,l = 1,\dots,m;$$

*и пусть подмножество* $Q_0 \subset Q$ *задается системой уравнений*

$$f_1(\theta) = 0,\ \dots,\ f_s(\theta) = 0, \quad \text{где} \quad f_1,\dots,f_s \in C^1(Q), \tag{19.3}$$

*причем* $\operatorname{rank}\{f_1'(\theta),\dots,f_s'(\theta)\} = s$ *в каждой точке* $\theta \in Q_0$.

*Тогда при* $\theta \in Q_0$ *распределение статистики* $2\ln\Lambda(X)$ *сходится к распределению* $\chi^2_s$. *В частности, если множество* $Q_0$ *состоит из одного элемента* $\theta_0$, *то статистика* $2\ln\Lambda(X)$ *сходится по распределению к* $\chi^2_m$.

Наметим план доказательства этой теоремы для случая $Q_0 = \{\theta_0\}$. Оно базируется на том, что в условиях регулярности *R*1), *R*2), *R*3) оценка максимального правдоподобия $\hat\theta$ оказывается асимптотически нормально распределенной с математическим ожиданием $\theta_0$ и матрицей ковариаций $\mathcal{I}_N(\theta_0)^{-1}$ (мы это доказали в случае одномерного $\theta$, см. теорему 8.3). Сделаем в пространстве параметров $\mathbb{R}^m$ линейную замену переменных, после которой информационная матрица $\mathcal{I}(\theta_0)$ становится единичной. Тогда распределение нормированной разности $\xi = \sqrt{N}(\hat\theta - \theta_0)$ сходится к стандартному нормальному закону (с нулевым средним и единичной матрицей ковариаций). Значит, координаты вектора $\xi$ в пределе становятся независимыми случайными величинами с распределением $\mathcal{N}(0,1)$. Отсюда вытекает, что распределение случайной величины $\xi^2 = N(\hat\theta - \theta_0)^2$ сходится к распределению $\chi^2_m$.

С другой стороны, раскладывая по формуле Тейлора функцию правдоподобия $l(\theta) = \ln p_\theta(X)$ в точке $\theta = \hat\theta$, получаем

$$\ln\Lambda(X) = \ln p_{\hat\theta}(X) - \ln p_{\theta_0}(X) = -\frac{1}{2}\frac{d^2 \ln p_\theta(X)}{d\theta^2}\big(\theta_0 - \hat\theta\big)^2, \qquad \theta \in \big[\hat\theta, \theta_0\big].$$

В силу закона больших чисел и сильной состоятельности оценки $\hat\theta$

$$\frac{1}{N}\frac{d^2 \ln p_\theta(X)}{d\theta^2} = \frac{1}{N}\sum_{i=1}^N \frac{d^2 \ln p_\theta(x_i)}{d\theta^2} \xrightarrow{\text{п. н.}} \mathsf{E}_{\theta_0}\left\{\frac{d^2 \ln p_\theta(x)}{d\theta^2}\bigg|_{\theta=\theta_0}\right\} = -\mathcal{I}(\theta_0) = -I.$$

Поэтому $2\ln\Lambda(X) \sim N\big(\theta_0 - \hat\theta\big)^2$ сходится по распределению к $\chi^2_m$.

В случае, когда множество $Q_0$ задается системой уравнений (19.3), статистика $\ln\Lambda(X)$ представляется в виде разности $\ln p_{\hat\theta}(X) - \ln p_{\bar\theta}(X) \sim N\big(\bar\theta - \hat\theta\big)^2$, где $\bar\theta$ — точка максимума функции $l(\theta) = p_\theta(X)$ на множестве $Q_0$. При этом оказывается, что вектор $\hat\theta - \bar\theta$ асимптотически совпадает с проекцией $\hat\theta - \theta_0$ на ортогональное дополнение к $Q_0$ в точке $\theta_0$. Поэтому $N\big(\bar\theta - \hat\theta\big)^2$ сходится по распределению к $\chi^2_s$.

# § 20. Байесовское решающее правило

Пусть параметр $\theta$ может принимать лишь конечное число значений $\theta_1, \ldots, \theta_k$, которым отвечают плотности вероятностей $p_1(x), \ldots, p_k(x)$ на пространстве $\mathbb{R}^n$ (по отношению к некоторой мере $\mu$). Гипотеза $H_i$ состоит в том, что $\theta = \theta_i \ \ (i = 1, \ldots, k)$.

Пусть еще заданы априорные вероятности $\pi_i = P(\theta_i)$, дающие в сумме единицу. Будем считать, что если верна гипотеза $H_i$, а выбор сделан в пользу гипотезы $H_j$, то приходится платить штраф $w(i, j)$. В этом случае $w(i, j)$ называется функцией потерь.

Мы хотим построить рандомизированное решающее правило $d(X)$ для проверки гипотез $H_j$ по выборке $X = (x_1, \ldots, x_N)$. Оно полностью определяется критическими функциями $\varphi_j(X)$, которые равны вероятностям выбора гипотез $H_j$ при данной выборке $X$. Очевидно, эти критические функции должны быть неотрицательными, а их сумма должна равняться единице.

*Функционалом риска* для рандомизированного решающего правила называется полное математическое ожидание потерь

$$r = \sum_{i=1}^{k} \pi_i \int_{\mathbb{R}^{nN}} p_i(X) \sum_{j=1}^{k} \varphi_j(X) w(i, j)\, \mu^N(dX).$$

Перегруппировав слагаемые, его можно записать так:

$$r = \int_{\mathbb{R}^{nN}} \sum_{j=1}^{k} \varphi_j(X) \left( \sum_{i=1}^{k} \pi_i p_i(X) w(i, j) \right) \mu^N(dX). \tag{20.1}$$

Принцип оптимальности Байеса предписывает выбирать решающее правило таким образом, чтобы отвечающий ему риск был минимален. Решающее правило, удовлетворяющее этому условию, называют *байесовским.*

Вычислим апостериорные вероятности параметров $\theta_1, \ldots, \theta_k$ при известной выборке $X$ с помощью формулы Байеса

$$P(\theta_i | X) = \frac{\pi_i p_i(X)}{p(X)}, \qquad p(X) = \sum_{j=1}^{k} \pi_j p_j(X).$$

Апостериорное среднее значение потерь при выборе гипотезы $H_j$ равно

$$r(H_j|X) = \sum_{i=1}^{k} P(\theta_i|X)w(i,j) = \frac{1}{p(X)} \sum_{i=1}^{k} \pi_i p_i(X) w(i,j).$$

Оказывается, чтобы получилось байесовское решающее правило, нужно каждой выборке $X$ ставить в соответствие ту гипотезу $H_j$, при которой апостериорное среднее значение потерь $r(H_j|X)$ минимально. Точнее, справедлива следующая теорема.

**Теорема 20.1.** *Рандомизированное решающее правило $d(X)$ является байесовским тогда и только тогда, когда оно при почти всех $X$ (в смысле меры $\mu^N$) выбирает лишь такие гипотезы $H_j$, для которых выполняется условие*

$$\sum_{i=1}^{k} \pi_i p_i(X) w(i,j) = \min_{1 \leqslant l \leqslant k} \sum_{i=1}^{k} \pi_i p_i(X) w(i,l). \tag{20.2}$$

*Это же верно и для нерандомизированных решающих правил.*

Доказательство. Функционал риска $r$ будет минимальным в том и только том случае, когда при почти всех $X$ минимальна подынтегральная функция в (20.1). А это равносильно тому, что вероятность $\varphi_j(X) = P\{d(X) = H_j\}$ отличается от нуля лишь для тех гипотез $H_j$, для которых выполняется условие (20.2). □

При $p(X) \neq 0$ условие (20.2) равносильно условию минимальности апостериорного риска $r(H_j|X)$. Если же $p(X) = 0$, то условие (20.2) выполняется для всех гипотез $H_j$, а апостериорные риски не определяются.

**Пример 1.** Построить байесовское решающее правило для функции потерь $w(i,j)$, которая равна единице при $i \neq j$ и нулю при $i = j$ (она называется (0–1)-функцией потерь).

*Решение.* Апостериорные вероятности $P(\theta_1|X)$, ..., $P(\theta_k|X)$ дают в сумме единицу. Поэтому апостериорное среднее значение потерь при выборе гипотезы $H_j$ равно

$$r(H_j|X) = \sum_{i \neq j} P(\theta_i|X) = 1 - P(\theta_j|X).$$

Это есть ни что иное, как апостериорная вероятность ошибки при выборе гипотезы $H_j$. Нам следует выбирать ту гипотезу, при которой эта вероятность минимальна, а вероятность $P(\theta_j|X) = \pi_i p_i(X)/p(X)$ максимальна.

**Пример 2.** Пусть заданы две плотности вероятности $p_0(x)$, $p_1(x)$ с априорными вероятностями $\pi_0$, $\pi_1$. Гипотеза $H_i$ состоит в том, что выборка $X$ взята из распределения с плотностью $p_i(x)$, где $i = 0, 1$. Если при истинной гипотезе $H_0$ принимается $H_1$, то штраф равен $w_0$, а если при истинной гипотезе $H_1$ принимается $H_0$, то штраф равен $w_1$. При выборе правильной гипотезы штраф нулевой. Найти байесовское решающее правило.

*Решение.* В данном случае апостериорные риски равны

$$r(H_0|X) = \frac{\pi_1 p_1(X) w_1}{p(X)}, \qquad r(H_1|X) = \frac{\pi_0 p_0(X) w_0}{p(X)}.$$

Нам следует выбирать ту гипотезу $H_j$, для которой апостериорный риск $r(H_j|X)$ минимален. Соответствующее решающее правило можно записать так:

$$d(X) = \begin{cases} H_0, & \text{если} \quad \dfrac{p_1(X)}{p_0(X)} \leqslant \dfrac{\pi_0 w_0}{\pi_1 w_1}, \\[2ex] H_1, & \text{если} \quad \dfrac{p_1(X)}{p_0(X)} > \dfrac{\pi_0 w_0}{\pi_1 w_1}. \end{cases}$$

## § 21. Последовательный анализ Вальда†

Чем больше объем выборки, тем достовернее статистические выводы, которые она позволяет сделать (меньше вариации точечных оценок, больше мощность решающих правил и т. п.) Для заранее заданного уровня достоверности всегда можно вычислить необходимый для его достижения объем выборки.

Идея последовательного анализа заключается в том, что для достижения необходимого уровня достоверности не нужно заранее фиксировать объем выборки, а его следует определять в ходе эксперимента в зависимости от ранее полученных выборочных значений. Таким образом, объем выборки становится случайным. Во многих задачах это позволяет существенно уменьшить среднее число наблюдений.

Последовательный анализ был основан в 1943 г. А. Вальдом. Он предложил использовать последовательный критерий отношения правдоподобия, который позволил примерно вдвое сократить среднее число наблюдений (при тех же вероятностях ошибок). Из-за этого открытие Вальда в годы Второй мировой войны было засекречено.

Рассмотрим следующий пример. Пусть на $\mathbb{R}^n$ имеются две всюду положительные плотности вероятностей $p_0(x)$ и $p_1(x)$ (по отношению к мере Лебега). Известно, что распределение случайной величины $x$ описывается одной из этих плотностей. Требуется определить, какой именно.

Будем рассматривать выборки $X_N = (x_1, \ldots, x_N)$ разных объемов. Мы хотим построить *последовательность* решающих правил $d_N(X_N)$, каждое из которых может принимать три значения: $H_0$, $H_1$ или $\bar{H}$. Если в какой-то момент времени оказывается, что $d_N(X_N) = H_0$, то принимается гипотеза $H_0$ (о том, что распределение случайных величин $x_t$ имеет плотность $p_0(x)$); если оказывается, что $d_N(X_N) = H_1$, то принимается гипотеза $H_1$ (о том, что распределение $x_t$ имеет плотность $p_1(x)$); наконец, в случае $d_N(X_N) = \bar{H}$ делается вывод о том, что полученной информации недостаточно для выбора определенной гипотезы, и принимается решение о добавлении к выборке очередного значения $x_{N+1}$. Описанная совокупность функций $d_N(X_N)$ называется *последовательным решающим правилом.*

Последовательный критерий отношения правдоподобия, который был предложен Вальдом, имеет вид

$$d_N(X_N) = \begin{cases} H_0, & \text{если } L(X_N) < A, \\ H_1, & \text{если } L(X_N) > B, \\ \bar{H}, & \text{если } A \leqslant L(X_N) \leqslant B, \end{cases} \tag{21.1}$$

где

$$L(X_N) = \frac{p_1(X_N)}{p_0(X_N)} = \prod_{t=1}^{N} \frac{p_1(x_t)}{p_0(x_t)}.$$

Введем обозначения:

$$z_t = \ln \frac{p_1(x_t)}{p_0(x_t)}, \qquad Z_N = \ln L(X_N) = z_1 + \ldots + z_N.$$

**Теорема 21.1.** *Если при каждой из гипотез $H_i$, $i = 0, 1$, существуют конечные математические ожидания $\mu_i = \mathsf{E}_i z_t$ и дисперсии $\sigma_i^2 = \mathsf{D}_i z_t$, то с вероятностью 1 критерий Вальда заканчивает свою работу за конечное число шагов.*

Доказательство. Предположим для определенности, что верна гипотеза $H_0$. Известно, что при всех положительных $x \neq 1$ имеет место неравенство $\ln x < x - 1$. Из него следует, что

$$\mu_0 = \mathsf{E}_0 z_t = \int_{\mathbb{R}^n} p_0(x_t) \ln \frac{p_1(x_t)}{p_0(x_t)}\, dx_t \;<\; \int_{\mathbb{R}^n} p_0(x_t) \left( \frac{p_1(x_t)}{p_0(x_t)} - 1 \right) dx_t = 0.$$

Очевидно, $\mathsf{E}_i Z_N = N\mu_i$ и $\mathsf{D}_i Z_N = N\sigma_i^2$. Поэтому

$$\begin{aligned} &P_0\{L(X_N) \geqslant A\} = P_0\{Z_N - N\mu_0 \geqslant \ln A - N\mu_0\} \leqslant \\ &\leqslant \mathsf{E}_0 \left\{ \frac{(Z_N - N\mu_0)^2}{(\ln A - N\mu_0)^2} \right\} \leqslant \frac{N\sigma_0^2}{(\ln A - N\mu_0)^2} \;\longrightarrow\; 0 \end{aligned} \tag{21.2}$$

(мы использовали то, что $\mu_0 < 0$ и, следовательно, $\ln A - N\mu_0 > 0$ при больших $N$).

Если $d_N(X_N) = \bar{H}$ при всех $N$, то $L(X_N) \geqslant A$ при всех $N$. Из (21.2) вытекает, что вероятность такого события равна нулю. $\square$

Пусть $m$ — момент остановки в критерии Вальда:

$$m = \min\{\, N \mid d_N(X_N) \neq \bar{H} \,\}.$$

По только что доказанной теореме $P_i\{m < \infty\} = 1$. Оценим вероятности ошибок первого и второго рода $\alpha$, $\beta$ для этого критерия. Очевидно,

$$\alpha = P_0\{d_m(X_m) = H_1\}, \qquad \beta = P_1\{d_m(X_m) = H_0\}.$$

Обозначим через $S_N^i$ множество всех выборок $X_N$, для которых $m = N$ и одновременно $d_N(X_N) = H_i$. Тогда в силу (21.1)

$$\begin{aligned} \beta = \sum_{N=1}^{\infty} P_1(S_N^0) = \sum_{N=1}^{\infty} \int\limits_{S_N^0} p_1(X_N)\, dX_N \leqslant \sum_{N=1}^{\infty} \int\limits_{S_N^0} A p_0(X_N)\, dX_N = \\ = A \sum_{N=1}^{\infty} P_0(S_N^0) = A(1 - \alpha), \end{aligned}$$

$$\alpha = \sum_{N=1}^{\infty} P_0(S_N^1) = \sum_{N=1}^{\infty} \int\limits_{S_N^1} p_0(X_N)\,dX_N \leqslant \sum_{N=1}^{\infty} \int\limits_{S_N^1} \frac{1}{B}\,p_1(X_N)\,dX_N =$$

$$= \frac{1}{B} \sum_{N=1}^{\infty} P_1(S_N^1) = \frac{1-\beta}{B}.$$

Таким образом, чтобы вероятности ошибок первого и второго рода не превосходили заданных уровней $\alpha_0$ и $\beta_0$, достаточно взять $A = \beta_0$ и $B = 1/\alpha_0$ в критерии (21.1).

Вычислим среднюю длину выборки, необходимую для принятия решения в критерии Вальда. Для этого нам потребуется следующая теорема.

**Теорема 21.2 (тождество Вальда).** *Если математические ожидания* $\mathsf{E}z_t$ *и* $\mathsf{E}m$ *конечны, то справедливо тождество* $\mathsf{E}Z_m = \mathsf{E}z_t \cdot \mathsf{E}m$ *(при каждой из гипотез* $H_0$, $H_1$*).*

Доказательство. Обозначим через $I_N$ характеристическую функцию множества исходов наблюдений, при которых $m = N$, а через $\chi_N$ — характеристическую функцию множества исходов, при которых $m > N$. Очевидно, $I_N = \chi_{N-1} - \chi_N$. Заметим еще, что случайная величина $\chi_{N-1}$ полностью определяется по выборке $X_{N-1}$, и потому не зависит от $z_N = \ln\bigl(p_1(x_N)/p_0(x_N)\bigr)$. Следовательно,

$$\mathsf{E}m = \sum_{N=1}^{\infty} NP\{m=N\} = \sum_{N=1}^{\infty} N\mathsf{E}I_N = \sum_{N=1}^{\infty} N\mathsf{E}\{\chi_{N-1} - \chi_N\} = \sum_{N=0}^{\infty} \mathsf{E}\chi_N,$$

$$\mathsf{E}Z_m = \sum_{N=1}^{\infty} \mathsf{E}\{Z_N I_N\} = \sum_{N=1}^{\infty}\sum_{t=1}^{N} \mathsf{E}\{z_t I_N\} = \sum_{t=1}^{\infty}\sum_{N=t}^{\infty} \mathsf{E}\{z_t I_N\} =$$

$$= \sum_{t=1}^{\infty} \mathsf{E}\{z_t \chi_{t-1}\} = \sum_{t=1}^{\infty} \mathsf{E}z_t\,\mathsf{E}\chi_{t-1} = \mathsf{E}z_t \sum_{N=0}^{\infty} \mathsf{E}\chi_N = \mathsf{E}z_t\,\mathsf{E}m. \quad \square$$

Из критерия Вальда вытекает, что гипотеза $H_0$ принимается при условии $Z_m < \ln A \leqslant Z_{m-1}$, а гипотеза $H_1$ принимается при условии $Z_m > \ln B \geqslant Z_{m-1}$. Отсюда следует, что при *истинной* гипотезе $H_0$ или $H_1$ выполняется соответственно одно из приближенных равенств

$$\mathsf{E}_0 Z_m \approx \ln A \cdot (1-\alpha) + \ln B \cdot \alpha, \qquad \mathsf{E}_1 Z_m \approx \ln A \cdot \beta + \ln B \cdot (1-\beta),$$

где погрешность не превосходит величины разности $Z_m - Z_{m-1} = z_m$. Значит,

$$\mathsf{E}_0 m = \frac{\mathsf{E}_0 Z_m}{\mathsf{E}_0 z_t} \approx \frac{\ln A \cdot (1-\alpha) + \ln B \cdot \alpha}{\mu_0} \leqslant \frac{\ln A}{\mu_0}, \tag{21.3}$$

$$\mathsf{E}_1 m = \frac{\mathsf{E}_1 Z_m}{\mathsf{E}_1 z_t} \approx \frac{\ln A \cdot \beta + \ln B \cdot (1-\beta)}{\mu_1} \leqslant \frac{\ln B}{\mu_1} \tag{21.4}$$

(правые части неравенств (21.3) и (21.4) положительны, потому что $A < 1 < B$ и $\mu_0 < 0 < \mu_1$).

Сравним эффективность критериев Вальда и Неймана — Пирсона на простейшем примере. Предположим, что на вещественной прямой заданы два нормальных распределения с одинаковыми дисперсиями $\sigma^2$ и разными математическими ожиданиями $\theta_0 < \theta_1$. Требуется различить гипотезы $\theta = \theta_0$ и $\theta = \theta_1$ так, чтобы вероятности ошибок I и II рода не превосходили $\varepsilon$.

В случае критерия Вальда достаточно взять $A = 1/B = \varepsilon$. При этом

$$\mu_0 = -\mu_1 = \int_{\mathbb{R}} p_0(x) \ln \frac{p_1(x)}{p_0(x)}\, dx =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{\mathbb{R}} e^{-(x-\theta_0)^2/2\sigma^2} \left( \frac{\theta_1 - \theta_0}{\sigma^2}\, x \,-\, \frac{\theta_1^2 - \theta_0^2}{2\sigma^2} \right) dx =$$

$$= \frac{\theta_1 - \theta_0}{\sigma^2}\, \theta_0 \,-\, \frac{\theta_1^2 - \theta_0^2}{2\sigma^2} = -\frac{(\theta_1 - \theta_0)^2}{2\sigma^2}.$$

Из последнего равенства и (21.3), (21.4) получаем

$$\mathsf{E}_0 m = \mathsf{E}_1 m \approx \frac{2\sigma^2 \ln(1/\varepsilon)(1-2\varepsilon)}{(\theta_1 - \theta_0)^2}. \tag{21.5}$$

В критерии Неймана — Пирсона с одинаковыми вероятностями ошибок первого и второго рода в силу симметрии следует выбирать гипотезу $\theta = \theta_0$ при условии $\bar{x} < (\theta_0 + \theta_1)/2$ и гипотезу $\theta = \theta_1$ при условии $\bar{x} > (\theta_0 + \theta_1)/2$. При этом вероятности ошибок равны

$$\alpha = \beta = P_0 \left\{ \bar{x} > \frac{\theta_0 + \theta_1}{2} \right\} = P_0 \left\{ \sqrt{N}\, \frac{\bar{x} - \theta_0}{\sigma} \;>\; \sqrt{N}\, \frac{\theta_1 - \theta_0}{2\sigma} \right\}.$$

Поскольку случайная величина $\sqrt{N}(\bar{x} - \theta_0)/\sigma$ имеет стандартное нормальное распределение, условие $\alpha = \beta \leqslant \varepsilon$ равносильно тому, что

$$\sqrt{N}\, \frac{\theta_1 - \theta_0}{2\sigma} \geqslant \Phi^{-1}(1-\varepsilon) \quad \Longleftrightarrow \quad N \geqslant \frac{4\sigma^2 \big(\Phi^{-1}(1-\varepsilon)\big)^2}{(\theta_1 - \theta_0)^2}. \tag{21.6}$$

Разделив равенство (21.5) на неравенство (21.6), получим отношение средней длины выборки в критерии Вальда к длине выборки в критерии Неймана — Пирсона, обеспечивающей ту же точность:

$$\frac{\mathsf{E}_0 m}{N} \;\leqslant\; \frac{\ln(1/\varepsilon)(1-2\varepsilon)}{2\big(\Phi^{-1}(1-\varepsilon)\big)^2}. \tag{21.7}$$

При $\varepsilon = 0{,}05$ получаем $\ln(1/\varepsilon) \approx 2{,}996$, $\Phi^{-1}(1-\varepsilon) \approx 1{,}645$, и вся правая часть (21.7) приближенно равна 0,49.

**Задача.** Докажите, что при $\varepsilon \to 0$ правая часть (21.7) стремится к $1/4$.

*Подсказка*: используйте асимптотику

$$1 - \Phi(x) \sim \frac{1}{\sqrt{2\pi}\, x}\, e^{-x^2/2} \quad \text{при} \quad x \to +\infty,$$

положив в ней $x = \Phi^{-1}(1-\varepsilon)$.

# Приложение. Необходимые сведения из теории вероятностей

## I. Случайные величины

*Борелевской* σ*-алгеброй* в $\mathbb{R}^n$ называется минимальная σ-алгебра, содержащая все открытые и замкнутые подмножества $\mathbb{R}^n$. Отображение $f : \mathbb{R}^n \to \mathbb{R}^m$ называют *борелевским,* если прообраз любого борелевского множества из $\mathbb{R}^m$ есть борелевское подмножество в $\mathbb{R}^n$.

*Вероятностным пространством* называется тройка $(\Omega, \mathfrak{A}, P)$, где $\Omega$ — произвольное множество, $\mathfrak{A}$ — некоторая σ-алгебра подмножеств в $\Omega$, а $P$ — вероятностная мера на $\mathfrak{A}$ (то есть такая неотрицательная мера, что $P(\Omega) = 1$). Множества $A \in \mathfrak{A}$ называются *событиями,* элементы $\omega \in \Omega$ называются *элементарными событиями* или *исходами,* а число $P(A) = P\{\omega \in A\}$ называется *вероятностью* события $A$.

Любая $\mathfrak{A}$-измеримая функция $\xi : \Omega \to \mathbb{R}$ (или измеримое отображение $\xi : \Omega \to \mathbb{R}^n$) называется *случайной величиной* (*или случайным вектором*). Измеримость $\xi$ означает, что прообраз любого борелевского множества из $\mathbb{R}$ (или $\mathbb{R}^n$) принадлежит σ-алгебре $\mathfrak{A}$.

В теории меры принят обычай пополнять исходную σ-алгебру $\mathfrak{A}$ до σ-алгебры измеримых множеств $\mathfrak{A}'$. Множество $A'$ называется *измеримым,* если существует такое множество $A \in \mathfrak{A}$, что симметрическая разность между $A'$ и $A$ содержится в некотором элементе $\mathfrak{A}$ нулевой меры. Очевидно, пополненная алгебра $\mathfrak{A}'$ зависит от исходной меры $P$ на $\mathfrak{A}$. В статистике объектом изучения являются *семейства* вероятностных мер на алгебре $\mathfrak{A}$, и поэтому ее *никогда не пополняют.* На самом деле это никак не ограничивает класс рассматриваемых случайных величин, потому что для каждой $\mathfrak{A}'$-измеримой функции существует отличающаяся от нее лишь на множестве меры нуль $\mathfrak{A}$-измеримая функция.

*Математическим ожиданием* случайной величины $\xi$ называется интеграл

$$\mathsf{E}\xi = \int_\Omega \xi \, dP$$

(если он существует). Множество всех случайных величин, определенных на вероятностном пространстве $(\Omega, \mathfrak{A}, P)$ и имеющих конечное математическое ожидание, совпадает с пространством интегрируемых функций $L^1(\Omega, \mathfrak{A}, P)$.

*Дисперсией* случайной величины $\xi$ называется число

$$\mathsf{D}\xi = \mathsf{E}\big\{(\xi - \mathsf{E}\xi)^2\big\} = \mathsf{E}\big\{\xi^2\big\} - (\mathsf{E}\xi)^2$$

(если оно определено). Случайные величины с конечной дисперсией образуют вещественное гильбертово пространство $H = L^2(\Omega, \mathfrak{A}, P)$. Скалярное произведение двух случайных величин $\xi, \eta \in H$ имеет вид

$$(\xi, \eta) = \int_\Omega \xi\eta \, dP = \mathsf{E}\{\xi\eta\}.$$

*Ковариацией* двух случайных величин $\xi, \eta \in H$ называется число

$$\mathrm{Cov}\{\xi, \eta\} = \mathsf{E}\big\{(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta)\big\}.$$

Ковариация между $\xi$ и $\eta$ линейно зависит от $\xi$ и от $\eta$. *Корреляцией* (или *коэффициентом корреляции*) между $\xi$ и $\eta$ называется число

$$\mathrm{Corr}\{\xi, \eta\} = \frac{\mathrm{Cov}\{\xi, \eta\}}{\sqrt{\mathsf{D}\xi}\,\sqrt{\mathsf{D}\eta}} = \frac{\mathsf{E}\big\{(\xi - \mathsf{E}\xi)(\eta - \mathsf{E}\eta)\big\}}{\sqrt{\mathsf{E}\{(\xi - \mathsf{E}\xi)^2\}}\,\sqrt{\mathsf{E}\{(\eta - \mathsf{E}\eta)^2\}}}$$

Из последнего равенства следует, что $\mathrm{Corr}\{\xi, \eta\}$ есть ни что иное, как косинус угла между векторами $\xi - \mathsf{E}\xi$ и $\eta - \mathsf{E}\eta$ в гильбертовом пространстве $H$. Если $\mathrm{Cov}\{\xi, \eta\} = 0$, то случайные величины $\xi$ и $\eta$ называют *некоррелированными*.

*Характеристической функцией* случайной величины $\xi$ называется функция

$$\varphi(\lambda) = \int_\Omega e^{i\lambda\xi} \, dP = \mathsf{E}\big\{e^{i\lambda\xi}\big\}, \qquad \lambda \in \mathbb{R},$$

где $i$ — мнимая единица. Она равномерно непрерывна на вещественной прямой, не превосходит по абсолютной величине единицы и удовлетворяет равенству $\varphi(0) = 1$. Если еще у $\xi$ есть математическое ожидание, то $\varphi'(0) = i\mathsf{E}\xi$. Для случайного вектора $\xi \in \mathbb{R}^n$ характеристическая функция определяется точно так же и обладает аналогичными свойствами, нужно только считать, что $\lambda \in \mathbb{R}^n$ и $\lambda\xi = \lambda_1\xi_1 + \ldots + \lambda_n\xi_n$.

Все случайные величины удовлетворяют *неравенству Чебышёва*

$$P\big\{|\xi - a| \leqslant \varepsilon\big\} = P\big\{\,\omega \in \Omega \;\big|\; |\xi(\omega) - a| \leqslant \varepsilon\,\big\} \leqslant \frac{\mathsf{E}\big\{|\xi - a|^k\big\}}{\varepsilon^k}$$

при любых положительных числах $k$, $\varepsilon$.

## II. Пространство реализаций

Пусть на вероятном пространстве $(\Omega, \mathfrak{A}, P)$ определен случайный вектор $\xi\colon \Omega \to \mathbb{R}^n$ (в частности, это может быть скалярная случайная величина $\xi\colon \Omega \to \mathbb{R}$). Обозначим буквой $\mathfrak{B}$ борелевскую $\sigma$-алгебру в $\mathbb{R}^n$. Случайный вектор $\xi$ автоматически порождает вероятностную меру $P_\xi$ на $\mathfrak{B}$ по следующему правилу:

$$P_\xi(B) = P\big\{\xi(\omega) \in B\big\} = P\big(\xi^{-1}(B)\big), \qquad B \in \mathfrak{B}.$$

Эта мера называется *распределением вероятностей* для случайного вектора $\xi$, а вероятностное пространство $(\mathbb{R}^n, \mathfrak{B}, P_\xi)$ называется его *пространством реализаций*.

Определим случайный вектор $x \in \mathbb{R}^n$ как *тождественное отображение* вероятностного пространства $(\mathbb{R}^n, \mathfrak{B}, P_\xi)$ в себя. Он будет иметь то же самое распределение вероятностей, что и $\xi$:

$$P\big\{\xi(\omega) \in B\big\} = P_\xi\{x \in B\}, \qquad B \in \mathfrak{B}. \tag{1}$$

Более того, для любой борелевской функции $f\colon \mathbb{R}^n \to \mathbb{R}$ будет

$$\mathsf{E}\big\{f(\xi)\big\} = \int_\Omega f\big(\xi(\omega)\big)\, dP(\omega) = \int_{\mathbb{R}^n} f(x)\, P_\xi(dx) = \mathsf{E}\big\{f(x)\big\}, \tag{2}$$

причем оба интеграла сходятся или расходятся одновременно. В частности, характеристическая функция

$$\varphi(\lambda) = \mathsf{E}\big\{e^{i\lambda\xi}\big\} = \int_{\mathbb{R}^n} e^{i\lambda x}\, P_\xi(dx) \tag{3}$$

есть ни что иное, как преобразование Фурье от распределения $P_\xi$.

Как правило, в теории вероятностей изучаются не все возможные события $A \in \mathfrak{A}$, а лишь те, которые имеют вид

$$\xi^{-1}(B) = \{\omega \in \Omega \mid \xi(\omega) \in B\}, \qquad B \in \mathfrak{B}.$$

Тождества (1) и (2) означают, что по отношению к таким событиям случайные векторы $\xi$ и $x$ ведут себя абсолютно одинаково. Поэтому можно совершенно безболезненно заменить $\xi$ на $x$, и вместо абстракт-

ного вероятностного пространства $(\Omega, \mathfrak{A}, P)$ рассматривать конкретное пространство реализаций $(\mathbb{R}^n, \mathfrak{B}, P_\xi)$. Попросту говоря, мы «забываем» про область определения случайного вектора $\xi$ и заменяем его на независимую переменную $x \in \mathbb{R}^n$ с распределением вероятностей $P_\xi$. Формулы (1) – (3) являются примерами такой замены.

Вышеописанная процедура перехода от абстрактного вероятностного пространства к пространству реализаций удобна во многих отношениях; обычно она применяется по умолчанию без дополнительных пояснений.

Если существует неотрицательная функция $p(x) \geqslant 0$, для которой выполняется тождество

$$P_\xi(B) = \int_B p(x)\,dx, \qquad B \in \mathfrak{B},$$

то она называется *плотностью вероятности* распределения $P_\xi$ (по отношению к мере Лебега $dx$). С помощью нее равенства (2) можно переписать так:

$$\mathsf{E}\big\{f(\xi)\big\} = \int_{\mathbb{R}^n} f(x)p(x)\,dx = \mathsf{E}\big\{f(x)\big\}.$$

Для каждой случайной величины $\xi\colon \Omega \to \mathbb{R}$ определяется *функция распределения*

$$F_\xi(x) = P\big\{\xi(\omega) \leqslant x\big\}.$$

Функция распределения всегда не убывает, полунепрерывна справа, стремится к единице при $x \to +\infty$ и к нулю при $x \to -\infty$. Для скалярной случайной величины $\xi$ равенства (2) принимают вид

$$\mathsf{E}\big\{f(\xi)\big\} = \int_{-\infty}^{\infty} f(x)\,dF_\xi(x) = \mathsf{E}\big\{f(x)\big\},$$

где интеграл следует понимать как интеграл Лебега — Сти́лтьеса. Если у распределения случайной величины $\xi$ имеется плотность $p(x)$, то

$$F_\xi(x) = \int_{-\infty}^{x} p(y)\,dy.$$

## III. Законы больших чисел

Случайные величины (случайные векторы) $\xi_1, \ldots, \xi_n$ называются *независимыми,* если для любых борелевских множеств $B_1, \ldots, B_n$ выполняется тождество

$$P\{\xi_1 \in B_1, \ldots, \xi_n \in B_n\} = \prod_{i=1}^{n} P\{\xi_i \in B_i\}.$$

Другими словами, распределение случайного вектора $(\xi_1, \ldots, \xi_n)$ является произведением распределений его компонент. Если распределения случайных величин $\xi_i$ имеют плотности $p_i(x_i)$, то распределение случайного вектора $(\xi_1, \ldots, \xi_n)$ тоже имеет плотность

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i).$$

В теории вероятностей всегда предполагается, что все рассматриваемые случайные величины являются функциями на одном и том же «универсальном» вероятностном пространстве. Поэтому вероятность одновременного выполнения нескольких событий определяется просто как вероятность их пересечения[1].

Математическое ожидание произведения независимых случайных величин равно произведению их математических ожиданий:

$$\mathsf{E}\{\xi_1 \xi_2 \ldots \xi_n\} = \prod_{i=1}^{n} \mathsf{E}\{\xi_i\},$$

а дисперсия суммы равна сумме дисперсий:

$$\mathsf{D}\{\xi_1 + \ldots + \xi_n\} = \mathsf{D}\xi_1 + \ldots + \mathsf{D}\xi_n.$$

Любые борелевские функции $f_1(\xi_1), \ldots, f_n(\xi_n)$ от независимых случайных величин $\xi_1, \ldots, \xi_n$ тоже независимы. Отсюда следует, что если компоненты случайного вектора $\xi = (\xi_1, \ldots, \xi_n)$ независимы, то его

[1] В квантовой механике случайные величины и вероятности имеют более сложную природу. Вероятность одновременного выполнения нескольких событий можно определить только тогда, когда эти события *коммутируют.* Для некоммутирующих событий понятие одновременного выполнения просто не имеет смысла.

характеристическая функция $\varphi_\xi(\lambda)$ равна произведению характеристических функций компонент:

$$\varphi_\xi(\lambda) = \mathsf{E}\left\{e^{\,i(\lambda_1\xi_1+\ldots+\lambda_n\xi_n)}\right\} = \prod_{j=1}^{n} \mathsf{E}\left\{e^{\,i\lambda_j\xi_j}\right\} = \prod_{j=1}^{n} \varphi_{\xi_j}(\lambda_j).$$

Пусть задана последовательность случайных величин $\xi_1$, $\xi_2$, $\xi_3$, ... с одинаковыми математическими ожиданиями $\mathsf{E}\xi_i = a$. Принято говорить, что эта последовательность удовлетворяет *закону больших чисел,* если последовательность средних арифметических $\zeta_n = (\xi_1+\ldots+\xi_n)/n$ сходится по вероятности к константе $a$. В явном виде это означает, что для любого $\varepsilon > 0$

$$P\left\{\left|\frac{\xi_1+\ldots+\xi_n}{n} - a\right| > \varepsilon\right\} \to 0 \quad \text{при} \quad n\to\infty.$$

Про ту же самую последовательность говорят, что она удовлетворяет *усиленному* закону больших чисел, если последовательность средних арифметических $\zeta_n$ сходится к $a$ почти наверное (с вероятностью 1). Из сходимости почти наверное вытекает сходимость по вероятности, поэтому усиленный закон больших чисел действительно является более сильным утверждением, чем просто закон больших чисел.

**Теорема (Колмогоров).** *Если случайные величины $\xi_1$, $\xi_2$, $\xi_3$, ... независимы, одинаково распределены и имеют математическое ожидание $a$, то для них выполняется усиленный закон больших чисел*[1].

## IV. Центральная предельная теорема

Пусть на метрическом пространстве $M$ задана последовательность борелевских вероятностных мер $P_n$. Про нее говорят, что она *сходится* (точнее, *слабо сходится*) к борелевской вероятностной мере $P$, если для любой непрерывной ограниченной функции $f\colon M\to\mathbb{R}$

$$\int_M f\,dP_n \;\longrightarrow\; \int_M f\,dP. \tag{4}$$

[1]На самом деле теорема Колмогорова — это частный случай эргодической теоремы Биркгофа, являющейся одним из фундаментальных результатов теории динамических систем.

Слабую сходимость вероятностных мер можно определять разными эквивалентными способами. Например, достаточно потребовать выполнение (4) для ограниченных непрерывных функций с ограниченным носителем, или для ограниченных функций, удовлетворяющих условию Липшица. Можно также сказать, что слабая сходимость определяется *слабой топологией*. Окрестность вероятностной меры $P$ в слабой топологии состоит из вероятностных мер $Q$, удовлетворяющих условиям

$$\left|\int_M f_i\,dQ - \int_M f_i\,dP\right| < \varepsilon, \qquad i = 1, \ldots, k,$$

где $f_1, \ldots, f_k$ — ограниченные непрерывные функции, а $\varepsilon$ — произвольное положительное число. Если метрическое пространство $M$ сепарабельно, то слабая топология на множестве борелевских вероятностных мер метризуема.

В том случае, когда множество $M$ совпадает с $\mathbb{R}^m$, слабая сходимость вероятностных мер $P_n \to P$ равносильна сходимости их характеристических функций $\varphi_n(\lambda) \to \varphi(\lambda)$ в каждой точке $\lambda \in \mathbb{R}^m$. Если же $M = \mathbb{R}$, то слабая сходимость вероятностных мер $P_n \to P$ равносильна сходимости их функций распределения $F_n(x) \to F(x)$ в каждой точке непрерывности функции $F(x) = P\big\{(-\infty, x]\big\}$.

Исключительно важную роль в теории вероятностей и математической статистике играет семейство *нормальных* распределений $\mathcal{N}(a, d)$ на вещественной прямой с плотностями вида

$$p(x) = \frac{1}{\sqrt{2\pi d}}\, e^{-(x-a)^2/2d}.$$

Нормальное распределение $\mathcal{N}(a, d)$ имеет математическое ожидание $a$, дисперсию $d$ и характеристическую функцию вида $\varphi(\lambda) = e^{ia\lambda - d\lambda^2/2}$. Нормальное распределение с нулевым математическим ожиданием и единичной дисперсией называется *стандартным*. Ему отвечает функция распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\,dy.$$

Если случайные величины $\xi_1, \ldots, \xi_n$ независимы и нормально распределены, то любая их линейная комбинация тоже нормально распределена, причем

$$\mathsf{E}\{\alpha_1\xi_1 + \ldots + \alpha_n\xi_n\} = \alpha_1\mathsf{E}\xi_1 + \ldots + \alpha_n\mathsf{E}\xi_n, \tag{5}$$

$$\mathsf{D}\{\alpha_1\xi_1 + \ldots + \alpha_n\xi_n\} = \alpha_1^2\mathsf{D}\xi_1 + \ldots + \alpha_n^2\mathsf{D}\xi_n. \tag{6}$$

**Центральная предельная теорема.** *Если случайные величины* $\xi_1$, $\xi_2$, $\xi_3$, ... *независимы, одинаково распределены, имеют математическое ожидание* $a$ *и дисперсию* $d$, *то*

$$P\left\{\frac{\xi_1+\ldots+\xi_n-na}{\sqrt{nd}}\leqslant z\right\}\ \longrightarrow\ \Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-x^2/2}\,dx$$

*для всех* $z\in\mathbb{R}$. *Другими словами, распределение случайной величины*

$$\zeta_n=\frac{\xi_1+\ldots+\xi_n-na}{\sqrt{nd}}$$

*слабо сходится к стандартному нормальному распределению* $\mathcal{N}(0,1)$.

Аналогичная теорема справедлива для последовательности независимых одинаково распределенных случайных векторов. Кроме того, имеются многочисленные обобщения центральной предельной теоремы на последовательности зависимых и разно распределенных случайных величин, но мы не будем на них останавливаться.

## V. Теорема Радона — Никóдима

Пусть в множестве $\Omega$ имеется некоторая σ-алгебра $\mathfrak{A}$, на которой задана σ-конечная мера $\mu$. Функция $\nu:\mathfrak{A}\to\mathbb{R}$ называется *зарядом* (или аддитивной функцией множества), если

$$\nu\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\nu(A_i)$$

для любых непересекающихся множеств $A_i\in\mathfrak{A}$. Заряд $\nu$ называется *абсолютно непрерывным* (относительно $\mu$), если из равенства $\mu(A)=0$ вытекает $\nu(A)=0$.

**Теорема (Радон — Никóдим).** *Если заряд* $\nu\colon\mathfrak{A}\to\mathbb{R}$ *абсолютно непрерывен, то он представляется в виде*

$$\nu(A)=\int_A f\,d\mu,\qquad A\in\mathfrak{A},$$

*где* $f$ — *некоторая* $\mathfrak{A}$*-измеримая интегрируемая функция.*

Функция $f$ из этой теоремы называется производной Радона — Никóдима $d\nu/d\mu$ или плотностью заряда $\nu$. Она определена с точностью до множества меры нуль.

## VI. Стандартные распределения вероятностей

Приведем перечень наиболее часто используемых распределений на вещественной прямой.

А) *Равномерное* распределение на отрезке $[a,b]$ имеет плотность

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & \text{если } x\in[a,b],\\[2ex] 0, & \text{если } x\notin[a,b],\end{cases}$$

математическое ожидание $(b+a)/2$ и дисперсию $(b-a)^2/12$.

Б) *Нормальное* распределение $\mathcal{N}(a,d)$ имеет плотность

$$p(x)=\frac{1}{\sqrt{2\pi d}}\,e^{-(x-a)^2/2d},$$

математическое ожидание $a$, дисперсию $d$ и характеристическую функцию $e^{ia\lambda-d\lambda^2/2}$. Нормальное распределение $\mathcal{N}(0,1)$ называется *стандартным*. Ему отвечает функция распределения

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-y^2/2}\,dy.$$

В) *Экспоненциальное* распределение с параметром $\lambda>0$ имеет плотность

$$p(x)=\begin{cases}\lambda e^{-\lambda x} & \text{при } x\geqslant 0,\\ 0 & \text{при } x<0,\end{cases}$$

функцию распределения $F(x)=1-e^{-\lambda x}$ (при $x\geqslant 0$), математическое ожидание $\lambda^{-1}$ и дисперсию $\lambda^{-2}$.

Г) *Гамма*-распределение с параметрами $\lambda, n>0$ имеет плотность

$$p(x)=\begin{cases}\dfrac{\lambda^n x^{n-1}e^{-\lambda x}}{\Gamma(n)} & \text{при } x>0,\\[2ex] 0 & \text{при } x\leqslant 0,\end{cases}$$

математическое ожидание $n/\lambda$ и дисперсию $n/\lambda^2$. Чаще всего встречаются гамма-распределения с $\lambda=1$ и натуральными $n$.

Д) Пусть $\xi_1, \ldots, \xi_n$ — независимые случайные величины с распределением $\mathcal{N}(0,1)$. Тогда случайная величина $\chi^2_n = \xi_1^2 + \ldots + \xi_n^2$ имеет распределение «хи-квадрат» с $n$ степенями свободы. Его математическое ожидание равно $n$, а дисперсия $2n$.

Е) Пусть $\xi_0, \xi_1, \ldots, \xi_n$ — независимые случайные величины со стандартным нормальным распределением $\mathcal{N}(0,1)$. Тогда распределение случайной величины

$$t_n = \sqrt{n}\,\frac{\xi_0}{\sqrt{\xi_1^2 + \ldots + \xi_n^2}}$$

называется *распределением Стьюдента* с $n$ степенями свободы. При $n \to \infty$ оно сходится к стандартному нормальному распределению.

Ж) Пусть $\chi^2_n$ и $\chi^2_m$ — две независимые случайные величины, имеющие распределения «хи-квадрат» с $n$ и $m$ степенями свободы соответственно. Тогда распределение случайной величины

$$F_{n,m} = \frac{m}{n}\frac{\chi^2_n}{\chi^2_m}$$

называется *распределением Фишера* с $n$ и $m$ степенями свободы.

З) *Распределение Бернулли* с вероятностью успеха $p$ сосредоточено в точках 0 и 1, причем $P(1) = p$ и $P(0) = 1 - p$. Математическое ожидание распределения Бернулли равно $p$, а дисперсия $p(1-p)$.

И) *Биномиальное распределение* с вероятностью успеха $p$ в $n$ испытаниях сосредоточено в точках $i = 0, 1, \ldots, n$, имеющих вероятности $P(i) = C_n^i p^i (1-p)^{n-i}$. Математическое ожидание биномиального распределения равно $np$, а дисперсия $np(1-p)$.

К) *Геометрическое распределение* сосредоточено в точках $i = 0, 1, 2, \ldots$, имеющих вероятности $P(i) = p(1-p)^i$. Его математическое ожидание равно $(1-p)/p$, а дисперсия $(1-p)/p^2$.

Л) *Распределение Пуассона* сосредоточено в точках $i = 0, 1, 2, \ldots$, которые имеют вероятности $P(i) = e^{-\lambda}\lambda^i/i!$ (где $\lambda$ — положительный параметр). Его математическое ожидание и дисперсия равны $\lambda$.

# Литература


1. *Боровков, А. А.* Математическая статистика / А. А. Боровков. М.: Наука, 1984. 472 с.

2. *Лагутин, М. Б.* Наглядная математическая статистика / М. Б. Лагутин. М.: Бином, 2007. 472 с.

3. *Харин, Ю. С.* Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук. Минск: БГУ, 2005. 279 с.

# Предметный указатель


$p$-уровень, 17, 62

аддитивная функция множества, 84

борелевская σ-алгебра, 77
борелевское отображение, 77

вариационный ряд, 18
вариация оценки, 9
вероятностное пространство, 77
вероятность, 77
  – условная, 39
выборка, 4, 5
выборочная
  – дисперсия, 6, 12
  – дисперсия несмещенная, 8
  – медиана, 18
  – оценка, 11
  – функция распределения, 13
  – характеристическая функция, 12
выборочное среднее, 6, 12

гипотеза, 4, 56
  – альтернатива, 56
  – нулевая, 56
  – простая, 56
  – сложная, 56
гистограмма, 16

дисперсия, 77
доверительный интервал, 51
  – асимптотический, 52
  – центральный, 51

закон больших чисел, 82
  – усиленный, 82
заряд, 84
  – абсолютно непрерывный, 84

измеримое множество, 77
измеримое отображение, 77
информация, 22
исход, 77

квантиль, 16
ковариация, 78
количество информации, 22
корреляция, 78
коэффициент корреляции, 78
критерий, 56
  – $\chi^2$ Пирсона, 65
  – Вальда, 73
  – Колмогорова, 67
  – отношения правдоподобия, 68
  – согласия, 64
  – факторизации, 45, 49

математическое ожидание, 77
  – условное, 34, 36
матрица
  – вариаций, 9
  – Грама, 11
  – информационная, 25
  – ковариаций, 9
медиана, 16
  – выборочная, 18
мера
  – доминирующая, 22
  – считающая, 21
метод
  – максимального правдоподобия, 27
  – моментов, 19
  – обратной функции, 52
  – Стьюдента, 52

момент, 12, 19
– выборочный, 12
– обобщенный, 21
мощность решающего правила, 57

независимые случайные величины, 81
неравенство
– информации, 21, 25
– Йенсена, 36
– Рао — Крамéра, 23
– Чебышёва, 78
несмещенный тест, 57

отношение правдоподобия, 58, 67
оценивание
– непараметрическое, 7
– параметрическое, 7
оценка
– асимптотически несмещенная, 7
– асимптотически эффективная, 26
– байесовская, 42
– выборочная, 11
– максимального правдоподобия, 28
– несмещенная, 7
– по методу моментов, 20
– подстановочная, 11, 21
– сильно состоятельная, 7
– состоятельная, 7
– статистическая, 4, 6
– строго состоятельная, 7
– точечная, 50
– эффективная, 26
ошибка
– второго рода, 57
– первого рода, 56

плотность вероятности, 80
полная σ–алгебра, 77
полная вариация оценки, 9
полная дисперсия оценки, 9
последовательный анализ Вальда, 72
преобразование Фурье, 79
принцип неопределенности, 24
принцип оптимальности, 57
пространство реализаций, 79

распределение
– $\chi^2$, «хи-квадрат», 55, 65, 86
– апостериорное, 42
– априорное, 41
– Бернулли, 86
– биномиальное, 86
– выборочное, 11
– гамма, 85
– геометрическое, 86
– дискретное, 21
– непрерывное, 21
– нормальное, 83, 85
– Пуассона, 86
– равномерное, 85
– стандартное нормальное, 83, 85
– Стьюдента, 86
– условное, 37–41
– Фишера, 86
– экспоненциальное, 85
– эмпирическое, 11
распределение вероятностей, 5, 79
расстояние Колмогорова, 66
решающее правило, 56
– байесовское, 70
– Неймана — Пирсона, 57
– нерандомизированное, 56
– последовательное, 73
– рандомизированное, 56
– состоятельное, 57

слабая сходимость мер, 82
слабая топология, 83
случайная величина, 77
случайный вектор, 77
смещение оценки, 7
событие, 77
  – элементарное, 77
состоятельный тест, 57
статистика, 4, 5
  – $\chi^2$ Пирсона, 65
  – достаточная, 44
  – Колмогорова, 66
  – порядковая, 18
  – Стьюдента, 54
сходимость
  – по вероятности, 7, 82
  – почти наверное, 7, 82
  – слабая, 82

теорема
  – Гливенко — Кантелли, 15
  – Колмогорова, 82
  – Колмогорова — Блэкуэлла — Рао, 44
  – Радона — Никóдима, 84
  – центральная предельная, 84
тест, 56
тождество Вальда, 75

уравнение правдоподобия, 28
уровень значимости, 57
условия регулярности, 29, 68
условная вероятность, 39

формула Байеса, 39
функционал риска, 42, 70
функция
  – борелевская, 77
  – измеримая, 77
  – критическая, 56
  – потерь, 42, 44
  – правдоподобия (Фишера), 27
  – распределения, 13, 80
  – характеристическая, 12, 78
  – Хевисайда, 13

центральная предельная теорема, 84

штраф, 42

экспоненциальное семейство, 47
элементарное событие, 77

# Оглавление

**Предмет математической статистики** . . . . . 3

**Глава 1. Статистическое оценивание параметров** . . . . . 5

§ 1. Основные понятия статистического оценивания . . . . . 5
§ 2. Вариации и ковариации оценок . . . . . 8
§ 3. Выборочные оценки . . . . . 11
§ 4. Квантили и $p$-уровни . . . . . 16
§ 5. Метод моментов . . . . . 19
§ 6. Неравенство Рао — Крамéра . . . . . 21
§ 7. Эффективные оценки . . . . . 25
§ 8. Метод максимального правдоподобия . . . . . 27
§ 9. Условные математические ожидания . . . . . 33
§ 10. Условные распределения . . . . . 37
§ 11. Байесовские оценки . . . . . 41
§ 12. Достаточные статистики . . . . . 44
§ 13. Доказательство критерия факторизации . . . . . 49
§ 14. Доверительные интервалы . . . . . 50

**Глава 2. Статистическая проверка гипотез** . . . . . 56

§ 15. Основные понятия . . . . . 56
§ 16. Решающее правило Неймана — Пирсона . . . . . 58
§ 17. Проверка простой гипотезы против сложной альтернативы . . . 61
§ 18. Критерии согласия . . . . . 64
§ 19. Критерий отношения правдоподобия . . . . . 67
§ 20. Байесовское решающее правило . . . . . 70
§ 21. Последовательный анализ Вальда . . . . . 72

**Приложение. Необходимые сведения из теории вероятностей** . 77

I. Случайные величины . . . . . 77
II. Пространство реализаций . . . . . 79
III. Законы больших чисел . . . . . 81
IV. Центральная предельная теорема . . . . . 82
V. Теорема Радона — Никóдима . . . . . 84
VI. Стандартные распределения вероятностей . . . . . 85

**Литература** . . . . . 87

**Предметный указатель** . . . . . 88